COK- Ho 5933-9-HP947

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NUMBER 7
JULY 2000

9

CUTTA UNIVERSITY

page
365 *Evolution and the Problem of Other Minds* Elliott Sober

REVIEW ESSAYS
387 James M. Joyce: *The Foundations of Causal Decision Theory* Isaac Levi

BOOK REVIEWS
403 Nancy Cartwright: *The Dappled World: A Study of the Boundaries of Science* Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine

409 NEW BOOKS
417 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

KP2AK

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc.

ISSN 0022–362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVII, NO. 7, JULY 2000

EVOLUTION AND THE PROBLEM OF OTHER MINDS*

The following diagram illustrates two inference problems. First, there is the strictly third-person behavior-to-mind problem[1] in which I observe your behavior and infer that you occupy some mental state. Second, there is the Self-to-Other problem, in which I notice that I always or usually occupy some mental state when I behave in a particular way; then, when I observe you produce the same behavior, I infer that you occupy the same mental state. This second inference is the subject of the traditional philosophical problem of other minds.

How are these inferences related? Notice that the inputs to behavior-to-mind inference are a subset of the inputs to Self-to-Other inference. In the first, I consider the behavior of the other individual; in the second, I consider that behavior, as well as my own behavior and mental state. This suggests that, if Self-to-Other inferences cannot be drawn because the evidence available is too meager, the same will be true of behavior-to-mind inferences as well.

I see no reason to think that a purely third-person scientific psychology is impossible. Of course, if one is a skeptic about *all* nondeductive inference, that skepticism will infect the subject matter of psychology. And if one thinks that scientific inference can *never*

---

*My thanks to Colin Allen, Martin Barrett, Marc Bekoff, Tom Bontly, Nancy Cartwright, James Crow, Frans De Waal, Ellery Eells, Mehmet Elgin, Berent Enç, Branden Fitelson, Peter Godfrey-Smith, Daniel Hausman, George Kampis, Richard Lewontin, Barry Loewer, David Papineau, Daniel Povinelli, Larry Shapiro, and Alan Sidelle for useful comments. I also have benefitted from discussing this paper at London School of Economics and Political Science, at University of Illinois/Chicago, at Caltech, at Eötös University, at the University of Vienna, and at Northern Illinois University.

[1] This terminology is from Larry Shapiro, "Presence of Mind," in Valerie Gray Hardcastle, ed., *Biology Meets Psychology: Constraints, Connections, and Conjectures* (Cambridge: MIT, 1999), pp. 133-50.

0022-362X/00/9707/365–86

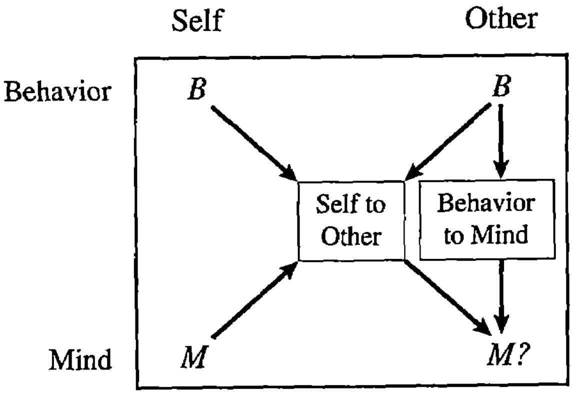


Figure 1

discriminate between empirically equivalent hypotheses, one also will hold that science is incapable of discriminating between such hypotheses when their subject matter is psychological. These are not special problems about psychology, however. What, then, becomes of the problem of other minds? If that problem concerns the tenability of Self-to-Other inference, it appears to be no problem, if science is able to draw behavior-to-mind inferences.

To see how the Self-to-Other problem can be detached from the problem of strictly third-person behavior-to-mind inference, we need to distinguish *absolute* from *incremental* versions of the Self-to-Other problem. The absolute problem concerns whether certain input information permits me to infer that the other person occupies mental state *M* rather than some alternative state *A*. As I have said, if information about the behavior of others permits me to infer that they are in mental state *M*, then it is hard to see why this inference should be *undermined* by adding the premise that I myself am in mental state *M* when I produce behavior *B*. The question remains, however, of whether first-person information makes a difference. To whatever degree third-person information provides an indication of whether the other person has *M* or *A*, does the addition of first-person information modify this assessment? This incremental version of the Self-to-Other problem is neutral on the question of whether third-person behavior-to-mind inference is possible. It is this incremental problem which I think forms the core of the problem of other minds, and this is the problem which I want to address here.[2]

I

Although the problem of other minds usually begins with an introspective grasp of one's own mental state, it can be detached from that setting and formulated more generally as a problem about "extrapolation." Thus, we might begin with the assumption that human beings produce certain behaviors because they occupy particular mental states and ask whether this licenses the conclusion that members of other species that exhibit the behavior do so for the same reason. None of us knows just by introspection, however, that *all* human beings who produce a given behavior do so because they occupy some particular mental state. In fact, this formulation of the problem of other minds, in which it is detached from the concept of introspection, is usually what leads philosophers to conclude that inferences about other minds from one's own case are weak. The fact that I own a purple bow tie should not lead me to conclude that you do, too. This point about bow ties is supposed to carry over to the fact that I have a mind and occupy various mental states. I know that I own a purple bow tie, but not by introspection.[3]

Discussion of the problem of other minds in philosophy seems to have died down (if not out) around thirty years ago.[4] Before then, it was discussed as an instance of "analogical" or "inductive" reasoning and the standard objection was that an inference about others based on your own situation is an extrapolation from too small a sample. This problem does not disappear merely by thinking of introspective experience as furnishing you with thousands of data points. The fact remains that they all were drawn from the same urn—your own mind. How can sampling from one urn help you infer the composition of another?

---

[2] There is a further reason for focusing on the incremental problem; it permits us to investigate the problem of other minds without formulating it as a problem about *acceptance*. The lottery paradox shows how difficult it is to say how much evidence in favor of a hypothesis is needed for one to believe the hypothesis. On this, see Henry Kyburg, *Probability and the Logic of Rational Belief* (Middletown, CT: Wesleyan, 1961).

[3] The fact that the problem of other minds can be detached from the idea of introspection does not entail that introspection has no special epistemic standing. Even if I can know things about my mind by pathways not open to you, the question remains of how I should extrapolate this self-knowledge to others. I should mention that I see no reason to think that the semantics of mentalistic terms entails that this epistemological problem cannot be posed intelligibly. Even if I fix the meaning of the word 'pain' by applying it to various experiences that I have had to date, it remains to be said whether the term also applies to others (or to myself at later dates).

[4] Perhaps one reason the problem largely lapsed from philosophical discussion is that it had become so closely connected with questions about logical behaviorism. When logical behaviorism went out of fashion, so did the problem of other minds. Behaviorism was replaced by a version of mentalism that emphasized the idea that third-person mentalistic hypotheses are to be judged, like other scientific hypotheses, on the basis of their ability to explain and predict. In consequence, the Self-to-Other problem was replaced by the behavior-to-mind problem.

When the problem is formulated in this way, it becomes pretty clear that what is needed is some basic guidance about inductive inference. The *mental content* of Self-to-Other inference is not what makes it problematic, but the fact that it involves *extrapolation*. Some extrapolations make sense while others do not. It seems sensible to say that thirst makes other people drink water, based on the fact that this is usually why I drink. Yet it seems silly to say that other folks walk down State Street at lunchtime because they crave spicy food, based just on the fact that this is what sets me strolling. By the same token, it seems sensible to attribute belly buttons to others, based on my own navel gazing. Yet it seems silly to universalize the fact that I happen to own a purple bow tie. What gives?

The first step toward answering this question started to emerge in the 1960s, not in philosophy of mind, but in philosophy of science. There is a general point about confirmation that we need to take to heart: observations provide evidence for or against a hypothesis only in the context of a set of background assumptions. If the observations do not deductively entail that the hypothesis of interest is true, or that it is false, then there is no saying whether the observations confirm or disconfirm, until further assumptions are put on the table. This may sound like the Duhem/Quine thesis, but that way of thinking about the present point is somewhat misleading, since Pierre Duhem[5] and W. V. Quine[6] discussed deductive, not probabilistic, connections of hypotheses to observations. If we want a person to pin this thesis to, it should be I. J. Good.[7] Good made this point forcefully in connection with Carl Hempel's[8] formulation of the ravens paradox. Hempel thought it was clear that black ravens and white shoes both confirm the generalization that all ravens are black. The question that interested him was why one should think that black ravens provide stronger confirmation than white shoes. Good responded by showing that empirical background knowledge can have the consequence that black ravens actually *dis*confirm the generalization. Hempel replied by granting that one could have special information that would undercut the assumption that black ravens and white shoes both confirm. He thought that in the absence of such information, however, it was a matter of logic that black ravens and white shoes are confirmatory. For this reason, Hempel asked the reader to indulge in a "methodological fiction." We are to imagine that we know nothing at all about the world, but are able to ascertain of individual objects what colors they have and whether they are ravens. We then are presented with a black raven and a white shoe; logic alone is supposed to tell us that these objects are confirming instances. Good's response was that a person in the circumstance described would not be able to say anything about the evidential meaning of the observations. I think that subsequent work on the concept of evidence, both in philosophy and in statistics, has made it abundantly clear that Good was right and Hempel was wrong. Confirmation is not a two-place relationship between observations and a hypothesis; it is a three-place relation between observations, a hypothesis, and a background theory.[9] Black ravens confirm the generalization that all ravens are black, given some background assumptions, but fail to do so, given others. And if no background assumptions can be brought to bear, the only thing one can say is: *out of nothing, nothing comes.*

---

[5] In *The Aim and Structure of Physical Theory* (Princeton: University Press, 1914/1954).

[6] In "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 20-46

[7] In "The White Shoe Is a Red Herring," *British Journal for the Philosophy of Science*, XVII (1967): 322, and in "The White Shoe Qua Herring Is Pink," *British Journal for the Philosophy of Science*, XIX (1968): 156-57.

[8] "Studies in the Logic of Confirmation," in *Aspects of Scientific Explanation and Other Essays* (New York: Free Press, 1965), and "The White Shoe: No Red Herring," *British Journal for the Philosophy of Science*, XVIII (1967): 239-40.

If we apply this lesson to the problem of other minds, we obtain the following result: the fact that I usually or always have mental property M when I perform behavior B is, by itself, *no reason at all* to think that you usually or always have M when you perform B. If that sounds like skepticism, so be it. The inference from Self to Other can make sense, however, once additional background assumptions are stated. A nonskeptical solution to the problem of other minds, therefore, must identify plausible further assumptions that bridge the inferential gap between Self and Other.[10]

---

[9] In fact, there are many important circumstances in which the evidence relation must be four-placed—a set of observations favors one hypothesis over another, relative to a set of background assumptions. This is the proper format for likelihood inference; see Richard M. Royall, *Statistical Evidence: A Likelihood Paradigm* (Boca Raton: Chapman and Hall/CRC, 1997), and my "Testability," *Proceedings and Addresses of the American Philosophical Association*, LXXIII (1999): 47-76; the latter is also available at the following URL—http://philosophy.wisc.edu/sober.

[10] I am not going to discuss in this paper what it means for two individuals to exhibit "the same behavior," but I shall make two comments. First, there is no requirement that they exhibit the same muscular movements; on this point, see Berent Enç, "Units of Behavior," *Philosophy of Science*, LXII (1995): 523-42. Second, however behaviors are individuated, it is important that one be able to say that two individuals share a behavioral trait without already knowing what the proximate

II

The problem of other minds has been and continues to be important in psychology (actually, in comparative psychology), except that there it is formulated in the first-person plural. Suppose that when human beings perform behavior *B*, we usually or always do so because we have mental property *M*. When we observe behavior *B* in another species, should we take the human case to count as evidence that this species also has mental property *M*? For this problem to be nontrivial, we assume that there is at least one alternative internal mechanism, *A*, which also could lead organisms to produce the behavior. Is the fact that humans have *M* evidence that this other species has *M* rather than *A*?

Discussion of this problem in comparative psychology has long been dominated by the fear of naive anthropomorphism.[11] This attitude was crystalized by C. Lloyd Morgan,[12] who suggested that, if we can explain a nonhuman organism's behavior by attributing to it a "higher" mental faculty, or by assigning it a "lower" mental faculty, then we should prefer the latter explanation. Morgan's successors embraced this "canon" because of its prophylactic qualities—it reduces the chance of a certain type of error. It is important to recognize, however, that there are *two* types of error that might occur in this situation:

	O lacks *M*	*O* has *M*
Deny that *O* has *M*		type-2 error
Affirm that *O* has *M*	type-1 error	

Morgan's canon does reduce the chance of type-1 error, but that is not enough to justify the canon. By the same token, a principle that encouraged anthropomorphism would reduce the chance of type-2 error, but that would not justify this liberal principle, either.

---

mechanism is (cognitive or otherwise) that leads them to produce the behavior; otherwise the inference problem here addressed could not get off the ground; see my "Black Box Inference: When Should an Intervening Variable Be Postulated?" *British Journal for the Philosophy of Science*, XLIX (1998): 469-98.

[11] See Daniel C. Dennett, "Intentional Systems in Cognitive Psychology: The 'Panglossian Paradigm' Defended," in *The Intentional Stance* (Cambridge: MIT, 1989), pp. 237-68, and Colin Allen and Marc Bekoff, *Species of Mind: The Philosophy and Biology of Cognitive Ethology* (Cambridge: MIT, 1997).

[12] In *An Introduction to Comparative Psychology* (London: Walter Scott, 1903).

Morgan tried to give a deeper defense of his canon; he thought he could justify it on the basis of Darwin's theory of evolution. Although Morgan's interesting argument does not work,[13] it is noteworthy that Morgan explicitly rejects what has become a fairly standard view of his principle—that it is a version of the principle of parsimony. Morgan thought that the simplest hypothesis about nonhuman organisms is that they are just like us. Parsimony favors anthropomorphism; the point of the canon is to counteract this tendency of thought. As we shall see, Morgan's conception of the relationship of his canon to the principle of parsimony was prescient.

I now want to leave Morgan's late-nineteenth century ideas about comparative psychology behind, and fast forward to the cladistic revolution in evolutionary biology that occurred in the 1970s and after. The point of interest here is the use of a principle of phylogenetic parsimony to infer phylogenetic relationships among species, based on data concerning their similarities and differences.[14] Although philosophers often say that parsimony is an ill-defined concept, its meaning in the context of the problem of phylogenetic inference is pretty clear. The hypotheses under consideration specify phylogenetic trees. The most parsimonious tree is the one that requires the smallest number of changes in character state in its interior to produce the observed distribution of characteristics across species at the tree's tips.

Consider the problem of inferring how sparrows, robins, and crocodiles are related to each other. Two hypotheses that might be considered are depicted in Figure 2. The (SR)C hypothesis says that sparrows and robins have a common ancestor that is not an ancestor of crocs; the S(RC) hypothesis says that it is robins and crocs that are more closely related to each other than either is to sparrows. Now consider an observation—sparrows and robins have wings, while crocodiles do not. Which phylogenetic hypothesis is better supported by this observation?

If winglessness is the ancestral condition, then the (SR)C hypothesis is more parsimonious. This hypothesis can explain the data about tip taxa by postulating a single change in character state on the branch with a slash through it. The S(RC) hypothesis, on the other hand, must postulate at least two changes in character state to explain the data. The principle of cladistic parsimony says that the observations favor (SR)C over S(RC). Notice that the parsimoniousness of a

---

[13] For discussion, see my "Morgan's Canon," in Colin Allen and Denise Dellarosa Cummins, eds., *The Evolution of Mind* (New York: Oxford, 1998), pp. 224-42.

[14] See, for example, Niles Eldredge and Joel Cracraft, *Phylogenetic Patterns and the Evolutionary Process* (New York: Columbia, 1980).

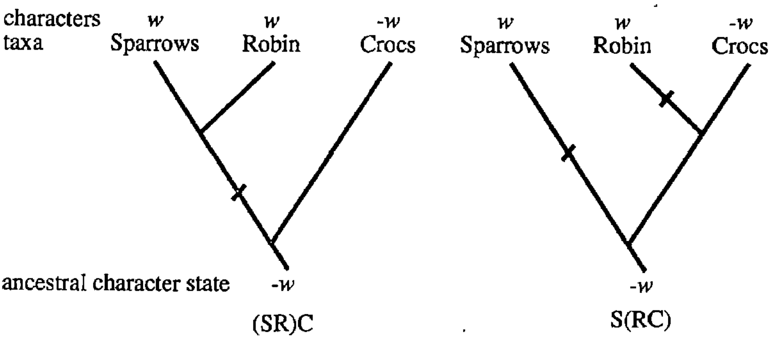


Figure 2

hypothesis is assessed not by seeing how many changes it says *actually* occurred, but by seeing what the *minimum* number of changes is that the hypothesis requires. (SR)C is more parsimonious because it entails a lower minimum.[15]

Although cladistic parsimony first attracted the attention of biologists because it helps one infer genealogical relationships, there is a second type of problem that parsimony allows one to address. If you use a set of traits to reconstruct the genealogy that connects several contemporaneous species (as in Figure 2), you can place a *new* set of traits on the tip species in that inferred tree and use parsimony to infer the character states of their ancestors. This inference is illustrated in Figure 3. Given the character states (1s and 0s) of tip species, the most parsimonious assignment of character states to interior nodes is the one shown.[16]

Parsimony is now regarded in evolutionary biology as a reasonable way to infer phylogenetic relationships. Whether it is the absolutely best method to use, in all circumstances, is rather more controversial. And the foundational assumptions that need to be in place for

---

[15] Cladistic parsimony does not entail that all similarities are evidence of common ancestry. For example, the two hypotheses depicted in Figure 2 would be equally parsimonious if wings were ancestral rather than derived. For discussion, see my *Reconstructing the Past: Parsimony, Evolution, and Inference* (Cambridge: MIT, 1988).

[16] Figure 3 illustrates a pattern of inference that is important in the study of human evolution. Why should features that are shared among current hunter-gatherer societies be thought to provide an indication of the ancestral human condition? After all, we cannot assume that hunter gatherers are "living fossils." The reason the inference makes sense is that current hunter gatherers are very distantly related to each other; this is what makes any similarities they may exhibit relevant to estimating the character state of the most recent common ancestor shared by all human lineages.

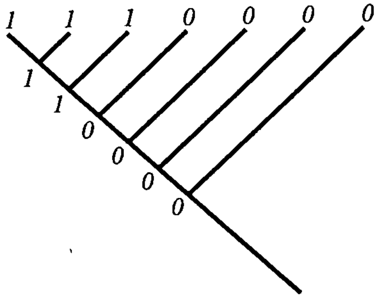


Figure 3

parsimony to make sense as an inferential criterion are also a matter of continuing investigation.

The reason I have explained the basic idea behind cladistic parsimony is that it applies to the problem of other minds, when Self and Other are genealogically related. Suppose both Self and Other are known to have behavioral characteristic *B*, and that Self is known to have mental characteristic *M*. The question is whether Other should be assigned *M* as well. As before, I shall assume that *M* is sufficient for *B*, but not necessary (an alternative internal mechanism, *A*, also could produce *B*). The two hypotheses we need to consider are depicted in Figure 4. If the root of the tree has the characteristic not-*B* (and so has neither-*M*-nor-*A*),[17] then the (Same) hypothesis is more parsimonious than the (Diff) hypothesis. It is consistent with (Same) that the postulated similarity linking Self and Other is a homology; it is possible that the most recent common ancestor of Self and Other had *M*, and that *M* was transmitted unchanged from this ancestor to the two descendants. The (Same) hypothesis, therefore, requires only a single change in character state, from neither-*M*-nor-*A* to *M*. In contrast, the (Diff) hypothesis requires at least two changes in character state.[18]

Frans De Waal[19] presents this cladistic argument in defense of the idea that parsimony favors anthropomorphism—we should prefer the

---

[17] We might have evidence that not-*B* is the ancestral character state by looking at a number of further individuals, besides Self and Other, who provide relevant "out-groups," as in Figure 3.

[18] Notice that, if no assumption is made about the character state at the root of the tree, (SAME) is more parsimonious. But if *A* is the ancestral character state, (SAME) and (DIFF) are equally parsimonious.

[19] In "Complementary Methods and Convergent Evidence in the Study of Primate Social Cognition," *Behaviour*, CXVIII (1991): 297-320, and in "Anthropomorphism and Anthropodenial: Consistency in Our Thinking about Humans and Other Animals," *Philosophical Topics* (forthcoming).

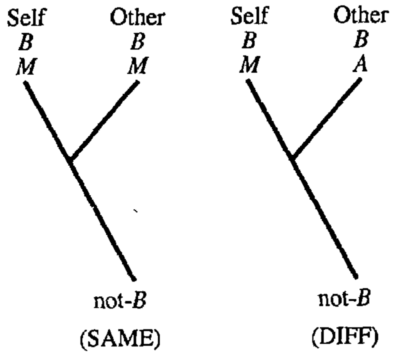


Figure 4

hypothesis that other species have the same mental characteristics that we have when they exhibit the same behavior.[20] This parsimony argument goes against Morgan's canon, just as Morgan foresaw. De Waal adds the reasonable proviso that the parsimony inference is strongest when Self and Other are closely related.

There is another proviso that needs to be added to this analysis, which is illustrated in Figure 5. As before, Self and Other are observed to have behavioral trait *B*. We know by assumption that Self has the mental trait *M*. The question, as before, is what we should infer about Other: Does it have *M* or *A*? The new wrinkle is that there are additional species depicted in the tree, ones that are known to lack *B*. The genealogical relationships that connect these further species to Self and Other entail that the most parsimonious hypothesis is that *B* evolved twice. It now makes no difference in parsimony whether one thinks that Other has *M* or *A*. What this shows is that parsimony favors anthropomorphism about mentalistic properties only when the behaviors in question are thought to be homologous.

This point has implications about a kind of question that often arises in connection with sociobiology. Parsimony does not oblige us to think that "slave-making" in social insects has the same psychosocial causes as slave-making in humans (or that rape in human beings has the same proximate mechanism as "rape" in ducks). The *behaviors* are not homologous, so there is no argument from parsimony for thinking that the same *proximate mechanisms* are at work. This is a point in favor of the parsimony analysis—cladistic parsimony explains why certain types of implausible inference really are implausi-

---

[20] See also my "Morgan's Canon"; and Shapiro, "Adapted Minds" (unpublished).

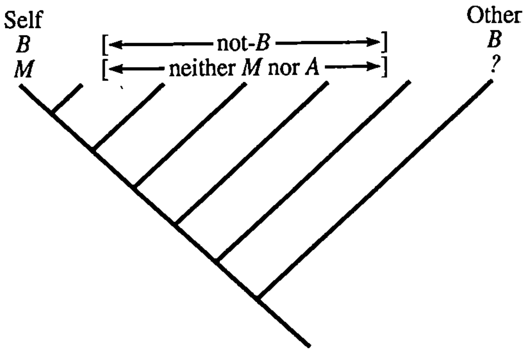


Figure 5

ble. *Parsimonious* anthropomorphism is not the same as *naive* anthropomorphism.

Just as extrapolation from Self to Other can be undermined by behavioral information about other species, extrapolation also can be undermined by neurophysiological information about Self and Other themselves. If Self has mental state *M* by virtue of being in physical state P1, what are we to make of the discovery that Other has physical state P2 (where P1 and P2 are mutually exclusive)? If we accept functionalism's assurance that mental state *M* is multiply realizable, this information does not entail that Other cannot be in mental state *M*; after all, P2 might just be a second supervenience base for *M*. On the other hand, P2 might be a supervenience base for alternative state *A*, and not for *M* at all. This inference problem is illustrated in Figure 6. For simplicity, let us suppose that P1 and P2 each suffice for *B*. P1 produces *B* by way of mental state *M*; it is not known at the outset whether P2 produces *B* by way of mental state *M*, or by way of alternative state *A*. We assume that the ancestor at the root of the tree did not exhibit the behavior in question, and that this ancestor had trait P0.

The conclusion we must reach is that the two hypotheses in Figure 6 are equally parsimonious. (Same) says that Self and Other both have *M* and that P1 and P2 are two of *M*'s supervenience bases. (Same) requires two changes to account for the characteristics at the tips; either P1 or P2 evolved long ago from an ancestor who had P0, and then either P1 changed to P2, or P2 changed to P1, in a subsequent branch of the tree. Of course, not-*B* changed to *B* in the same branch in which P0 changed to either P1 or P2, but this does not count as *two* changes, since the first entails the second. The (Diff) hypothesis also requires two changes if the root of the tree's having

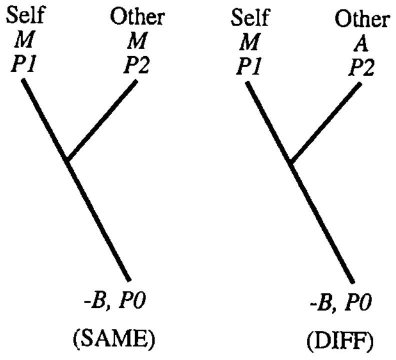


Figure 6

P0 is to be transformed into Self's having P1 and Other's having P2. Of course, (Diff) also will require changes from neither-*M*-nor-*A* at the root to *M* and *A* at the tips, but these do not count as changes *additional* to the ones that occur among P0, P1, and P2. Thus, the discovery of relevant neurophysiological differences can provide a context in which anthropomorphism is not sanctioned by parsimony considerations.

As a final exercise in understanding how the machinery of phylogenetic parsimony applies to the problem of extrapolating from Self to Other, let us consider the current controversy in cognitive ethology concerning whether chimps have a theory of mind.[21] It is assumed at the outset that chimps have what Daniel C. Dennett (*op. cit.*) terms first-order intentionality; they are able to formulate beliefs and desires about the extra-mental objects in their environment. The question under investigation is whether they, in addition, have second-order intentionality. Do they have the ability to formulate beliefs and desires about the mental states of Self and Other? Adult human beings have both first- and second-order intentionality. Do chimps have both, or do they have first-order intentionality only?

Figure 7 provides a cladistic representation of this question. I assume that the ancestral condition is the absence of both types of intentionality, and that second-order intentionality can evolve in a lineage only if first-order intentionality is already in place. The point to notice is that parsimony considerations do not discriminate between the two hypotheses. The (SAME) hypothesis and the (DIFF)

---

[21] See the essays collected in Peter Carruthers and Peter K. Smith, eds., *Theories of Theories of Mind* (New York: Cambridge, 1995), and C. M. Heyes, "Theory of Mind in Nonhuman Primates," *Behavioral and Brain Sciences*, XXI (1998): 101-48.

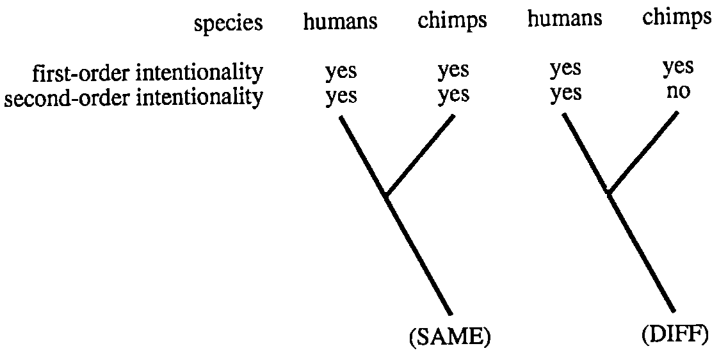


Figure 7

hypothesis both require two changes in the tree's interior—first- and second-order intentionality each must evolve at least once. The principle of parsimony in this instance tells us to be agnostic, and so disagrees with Morgan's canon, which tells us to prefer (DIFF).

Why does the problem formulated in Figure 7 lead to a different conclusion from the problem represented in Figure 4 in terms of the two states *M* and *A*? In Figure 4, we find that parsimony considerations favor extrapolation; in Figure 7, we find that extrapolating is no more and no less parsimonious than not extrapolating. Notice that Figure 4 mentions the behavior *B*, which internal mechanisms *M* and *A* are each able to produce. Figure 7, however, makes no mention of the behaviors that first- and second-order intentionality underwrite. This is the key to understanding why the analyses come out differently.

The point behind Figure 4 is this: if two species exhibit a homologous behavior, each of them must also possess an internal mechanism (*M* or *A*) for producing the behavior. It is more parsimonious to attribute the same mechanism to both Self and Other than to attribute different mechanisms to each. This pattern of argument can be applied to the problem depicted in Figure 7 by making explicit the behavioral consequences that first- and second-order intentionality are supposed to have. Suppose that first-order intentionality allows the organisms that have it to exhibit a range of behaviors B1. When a species that has first-order intentionality evolves second-order intentionality, this presumably augments the behaviors that organisms in the species are able to produce; the repertoire expands to

B1&B2.[22] If human beings exhibit B1 because they have first-order intentionality, we parsimoniously explain the fact that chimps also exhibit B1 (if they do) by saying that chimps have first-order intentionality. But there is no *additional* gain in parsimony to be had from attributing second-order intentionality to them as well, unless we also observe them producing the behaviors in B2.[23]

Cognitive ethologists are trying to find behaviors that chimps will produce if they have second-order intentionality but will not produce if they possess first-order intentionality only.[24] Identifying behaviors of this sort would permit an empirical test to be run concerning whether chimps have a theory of mind. It certainly is desirable that such behaviors should be found. Even if chimps exhibit B2, however, the question arises of why we should explain this by attributing second-order intentionality to them, rather than some alternative mechanism *A*. Of course, *A* will not be the trait of purely first-order intentionality, but presumably there are more than two options to be considered here. It is at this point that the principle of parsimony makes its entrance; if chimps exhibit behavior B2, this is better explained by the hypothesis that they have second-order intentionality than by the hypothesis that they have alternative mechanism *A*.

III

Why should we trust the parsimony arguments just described? Is parsimony an inferential end in itself or is there some deeper justification for taking parsimony seriously? In the present circumstance at least, there is no need to regard parsimony as something that we seek for its own sake.[25] My suggestion is that parsimony matters in problems of phylogenetic inference only to the extent that it reflects *likelihood.*[26] Here, I am using the term 'likelihood' in the technical sense introduced by R. A. Fisher. The likelihood of a hypothesis is the probability it confers on the observations, not the probability that the observations confer on the hypothesis. The likelihood of H, relative to the data, is Pr(Data | H), not Pr(H | Data).

---

[22] Some of the most interesting experiments on whether chimps have a theory of mind have been carried out by Daniel Povinelli; he looks for a B2 and finds that it is absent. I discuss his "knower/guesser" experiment in "Black Box Inference." More recently, Povinelli and Steve Gambrone have argued that "a novel psychological system for generating and sustaining higher-order representations, including the representation of other minds, may have evolved in the human lineage without *radically* altering our *basic* behavior patterns [italics mine]"—see their "Inferring Other Minds: Failure of the Argument by Analogy," *Philosophical Topics* (forthcoming). The words I have italicized in this quotation are important; the authors do not deny that there are behavioral differences between present day human beings and chimps that reflect the fact that the former have a theory of mind while the latter do not. Their suggestion is that the *initial appearance* of second-ordering intentionality may have had little or no immediate effect on behavior, but may have been a *building block* that allowed more substantial behavioral divergence to occur later.

[23] Suppose there is a behavior that human beings sometimes produce by using first-order intentionality only and sometimes by using second-order intentionality. What inference should we draw if we observe chimps producing this behavior? Again, there is no difference in parsimony between (SAME) and (DIFF).

[24] In fact, a behavioral test need not take this extreme form. It would suffice to find behaviors that are *probable* if subjects have a theory of mind but *improbable* if they have first-order intentionality only. The behavior need not be *impossible* in the absence of a theory of mind.

To see how the likelihood concept can be brought to bear on cladistic parsimony, consider Figure 2. It can be shown, given some minimal assumptions about the evolutionary process, that the data depicted in that figure are made more probable by the (RS)C hypothesis than they are by the R(SC) hypothesis. These assumptions are as follows:

> *Heritability*: character states of ancestors and descendants are positively correlated; that is, Pr(Descendant has a wing | Ancestor has a wing) > Pr(Descendant has a wing | Ancestor lacks a wing).
>
> *Chance*: all probabilities are strictly between 0 and 1.
>
> *Screening-off*: lineages evolve independently of each other, once they branch off from their most recent common ancestor.

What we have here, in its essentials, is a proof that Hans Reichenbach[27] gave in connection with his "principle of the common cause."

I think that most biologists would agree that these three assumptions hold pretty generally. The first of them does *not* say that descendants probably end up resembling their ancestors. The claim is *not* that stasis is more probable than change—that Pr(Descendant has a wing | Ancestor has a wing) > Pr(Descendant lacks a wing | Ancestor has a wing). Rather, the claim is that, if a descendant has a wing, that this result would have been more probable if its ancestor had had a wing than it would have been if its ancestor had lacked a wing.[28] The last assumption, it should be noted, is not exceptionless;

---

[25] The same point can be made in connection with the use of a parsimony criterion in problems of model selection, including curve-fitting problems; see Malcolm Forster and my "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, XLV (1994): 1-36.

[26] I defend this suggestion in my *Reconstructing the Past*.

[27] In *The Direction of Time* (Berkeley: California UP, 1956). For discussion, see chapter 3 of my *Reconstructing the Past*.

[28] The concept of heritability used here is closer to *narrow sense* heritability than to *broad sense* heritability; see D. S. Falconer and Trudy F. C. Mackay, *An Introduction to Quantitative Genetics* (London: Longmans, 1996, 4th ed.). It neither implies nor is implied by the idea of *genetic determination*. A trait can be heritable and still be

after all, there are ecological circumstances in which a lineage's evolving a trait influences the probability that other contemporaneous lineages will do the same. But even here, the assumption of independence is often true; and when it is false, it usually can be weakened without materially affecting the qualitative conclusions I want to draw.

Not only does likelihood provide a framework for understanding the role of parsimony considerations in phylogenetic inference; it also has implications about the Self and Other problem depicted in Figure 4. If traits M and A obey the assumptions listed, the following inequality is a consequence:

(P) $\Pr(\text{Self has } M \mid \text{Other has } M) > \Pr(\text{Self has } M \mid \text{Other has } A)$.

This proposition says that Self and Other are correlated (nonindependent) with respect to the traits M and A. It also says that there is a likelihood justification for anthropomorphism.[29] The observation that Self has M is rendered more probable by the hypothesis that Other has M than by the hypothesis that Other has A. Such differences in likelihood are generally taken to indicate a difference in support—the observation *favors* the first hypothesis over the second.[30]

The likelihood concept also throws light on De Waal's proviso—that parsimonious anthropomorphism is on firmer ground for our near relatives than it is for those individuals to whom we are related more distantly. We may translate this into the claim that the two

---

influenced by the environment. And a trait can be genetically determined and still not be heritable; the traits male (XY) and female (XX) provide examples.

[29] Hybrids aside, two species have a unique species that is their most recent common ancestor. Sexually reproducing individuals are not like this, however. Full-sibs have *two* parents as their most recent common ancestors. And first cousins usually overlap only partially in the set of ancestors they have two generations back; each has four grandparents, but (usually) only two of them are shared. How, then, does the Reichenbachian picture of common causes apply to human genealogies? The simplest way to connect them is to think of the common causes as sets of individuals, not as singletons. Thus, two full-sibs have the same parental *pair* as a common cause. And two first cousins can be thought of as tracing back to a set of six individuals; this is the set of all their grandparents, including the ones they do not share. Standard Mendelian genetics assures us that the state of this set screens off one cousin's genotype from the other's. Of course, this set is not limited to the two cousins' *common* ancestors. It can be shown that the common ancestors do screen off, however, since the ancestors in the set who are not shared influence one cousin's genotype but not the other's; see my and Martin Barrett's "Conjunctive Forks and Temporally Asymmetric Inference," *Australasian Journal of Philosophy*, LXX (1992): 1-23.

[30] See Royall; and my "Testability."

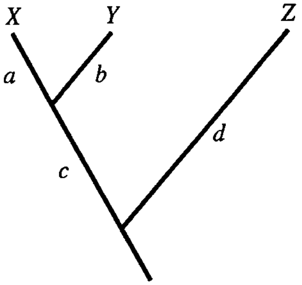


Figure 8

probabilities compared in proposition (P) are more different when Self and Other are closely related and become more similar as the relationship becomes more distant. This thesis is illustrated by Figure 8, in which X and Y are more closely related to each other than either is to Z. The lower-case letters in the tree's interior are path coefficients, which entail that the correlation of X and Y (r_{XY}) and the correlation of X and Z (r_{XZ}) have the values $r_{XY} = ab$ and $r_{XZ} = acd$. Notice that $r_{XY} > r_{XZ}$ if and only if $b > cd$. This inequality need not be true, but it does follow from an assumption that often figures in evolutionary discussions. This is the assumption of *uniform rates*—that a given evolutionary event has the same probability of occurring in different contemporaneous branches of a tree. In the present example, uniform rates means that $a = b$ and $d = ac$, which suffices to insure that X and Y are more strongly correlated than are X and Z. Thus, De Waal's proviso[31] is not true unconditionally; but when it is true, it need not be added as an independent constraint on parsimony arguments (which have no machinery for taking account of recency of divergence). The proviso flows from a likelihood analysis.

---

[31] I hope it is clear that this claim does not underwrite racist, nationalist, or species-ist conclusions. The argument does not provide a reason for denying that individuals outside one's own "group" have minds. For one thing, it is a mistake to think of one's self as belonging to just one group; each organism belongs to multiple, nested groups. For another, the argument presented here does not concern acceptance or rejection. And, finally, it is important to remember that Self-to-Other inference is not the only pathway by which we form beliefs about the internal states of others. There is, in addition, the possibility of strictly third-person behavior-to-mind inference. Much of our confidence concerning the mental states of others presumably rests on this second sort of inference. The point I am making about the *incremental* Self-to-Other problem is that the increment provided by knowledge of Self falls off as genealogical relatedness becomes more distant.

IV

Genealogical relatedness suffices to justify a likelihood inference from Self to Other, if the Reichenbachian assumptions that I described hold true. But is genealogical relatedness *necessary* for this extrapolation to make sense? Proposition (P) says that Self and Other are correlated with respect to trait *M*. What could induce this correlation? Reichenbach argued that whenever there is a correlation of two events, either the one causes the other, or the other causes the one, or the two trace back to a common cause. Considerations from quantum mechanics suggest that this is not always the case,[32] and doubts about Reichenbach's principle can arise from a purely classical point of view as well.[33] If we are not prepared to suppose that mentalistic correlations between Self and Other are brute facts, however, and if Self's having *M* does not causally influence whether Other has *M* (or vice versa), then Reichenbach's conclusion seems reasonable, if not apodictic; if Self and Other are correlated, this should be understood as arising from a common cause.

Genealogical relatedness is one type of common-cause structure. It can induce the correlation described in (P) by having ancestors transmit genes to their descendants, but there are alternatives that we need to recognize. To begin with, parents exert *non*genetic influences on their offspring through teaching and learning. For example, children have a higher probability of speaking Korean if their parents speak Korean than if their parents do not, but this is not because there is a gene for speaking Korean. And there are nongenetic connections between parents and offspring that do not involve learning, as when a mother transmits immunity to her children through her breast milk.

Correlations between Self and Other also can be induced by common causes when Self and Other are not genealogically related. If students resemble their teachers, then students of the same teacher will resemble each other. Here, learning does the work that genetic transmission is also able to do. Similarly, Self and Other can be correlated when they are influenced by a common environmental cause that requires no learning. For example, if influenza is spreading through one community, but not through another, then the fact

---

[32] See Bas van Fraassen, "The Charybdis of Realism: Epistemological Implications of Bell's Inequality," *Synthese*, LII (1982): 25-38.

[33] See van Fraassen, "Rational Belief and the Common Cause Principle," in Robert McLaughlin, ed., *What? Where? When? Why? Essays in Honor of Wesley Salmon* (Boston: Reidel, 1982), pp. 193-209; Nancy Cartwright, *Nature's Capacities and Their Measurement* (New York: Oxford, 1989); and my *Reconstructing the Past*, chapter 3.

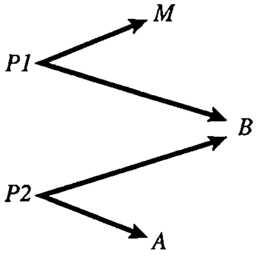


Figure 9

that I have the flu can be evidence that Other does, too, if the two of us live in the same community.

I list these alternatives, not because they apply with equal plausibility to the problem of other minds, but to give an indication of the range of alternatives which needs to be considered. Genealogical relatedness is only an example; the fundamental question is whether there are common causes impinging on Self and Other that induce the correlation described in (P). If there are, then there will be a likelihood justification for extrapolating from Self to Other.[34]

V

What, exactly, does cladistic parsimony and its likelihood analysis tell us about the problem of other minds? When I cry out, wince, and remove my body from an object inflicting tissue damage, this is (usually) because I am experiencing pain. When other organisms (human or not) produce the same set of behaviors, is this evidence that they feel pain? This hypothesis about Other *is* more parsimonious (if the behaviors are homologous and there are no known relevant neurophysiological differences), and it *is* more likely (if Reichenbachian assumptions about common causes apply). Does that completely solve the problem of other minds, or does there remain a residue of puzzlement?

One thing that is missing from this analysis is an answer to the question—*how much* evidence does the introspected state of Self provide about the conjectured state of Other? I have noted that the likelihoods in proposition (P) become more different as Self and

---

[34] Although I have described *M* and *A* as possible causes of the behavior *B*, this is not essential for the parsimony or likelihood arguments I have presented. Suppose that *M* and *A* are *epiphenomenal consequences* of the physical states P1 and P2, and are related to the behavior *B* as shown in Figure 9. If P1 suffices for *M* and *B* while P2 suffices for *B* and *A*, parsimony and likelihood are relevant to deciding whether *M* or *A* should be attributed to Other, given that Self has *M*.

Other become more closely related, but this comparative remark does not entail any quantitative benchmarks. It is left open whether the observation *strongly* favors one hypothesis over the other, or does so only *weakly*.

Another detail that I have not addressed is how probable it is that Other has *M*. If I have *M* when I produce behavior *B*, and Other exhibits *B*, is the probability greater than ½ that Other has M as well? The previous discussion helps answer that question, in that principle (P) is equivalent to the claim that Self's having *M* raises the probability that Other has *M*. Whatever the prior probability is that Other has *M*, the posterior probability is greater. Whether additional information can be provided that allows the value of that posterior probability to be estimated is a separate question.

A genealogical perspective on the problem of other minds helps clarify how that problem differs from the behavior-to-mind problem discussed at the outset. At first glance, it might appear that it does not matter to the problem of other minds whether the other individual considered is a human being, a dog, an extraterrestrial, or a computer. In all these cases, the question can be posed as to whether knowledge of one's own case permits extrapolation to another system that is behaving similarly. We have seen that these different formulations receive different answers. I share ancestors with other human beings, and I share other, more remote, ancestors with the nonhuman organisms found on earth. If there are creatures on other planets that evolved independently of life on earth, however, then I share no common ancestors with them. In this case, extrapolation from Self to Other does not have the genealogical justification I have described. I would go further and speculate that *it has no justification at all*. This does not mean that we should never attribute mental states to such creatures. What it does mean is that we must approach such questions as instances of the purely third-person behavior-to-mind inference problem. Similar remarks also may apply to computers. When they behave similarly to us (perhaps by passing an appropriate Turing test or by playing a competent game of chess), we may ask what causes them to do so. We have no ancestors in common with them; rather, we have constructed them so that they produce certain behaviors. Does this fact about the design of computers provide a reason to think that the proximate mechanisms behind those behaviors resemble those found in human beings? Arguably not. For extraterrestrials and (arguably) for computers, extrapolation from one's own case will not be justified. Indeed, this point applies to organisms with whom we do share ancestors, if the behaviors we have in common with them are not homologies (Figure 5). And even when Self and Other do exhibit a behavioral homology, if Self and Other are known to deploy different neural machinery for exhibiting that behavior, the extrapolation of *M* from Self to Other is undermined (Figure 6).

A probabilistic representation of the problem of other minds shows that the usual objection to extrapolating from Self to Other is in fact irrelevant, or, more charitably, that it rests on a factual assumption about the world that we have no reason to believe. The question is not whether introspected information about one's own mind provides lots of data or only a little. Rather, the issue is how strong the correlation is between Self and Other. Consider two urns that are filled with balls; each ball has a color, but the frequencies of different colors in the urns are unknown. The urns may be similar or identical in their compositions, or they may be very different. If I sample one ball from the first urn and find that it is green, does this provide substantial evidence concerning the second urn's composition? If I sample a thousand balls from the first urn, does this allow me to say any more about the second? Everything depends on how the two urns are related. If they are *independent*, then samples drawn from the first, whether they are small or large, provide no information about the second. But if they are not independent, then even a small sample from the first may be informative with respect to the second. There is nothing wrong with *asking* whether knowledge of Self supports a conclusion about Other. But the skeptical *assertion* that it does not involves a factual claim about the world. A claim of independence is no more a priori than a claim of correlation. If the relevant mental states of Self and Other are joint effects of a common cause (with the properties that Reichenbach described), then the skeptical assertion is false.

The argument I have presented is intended to show how certain *propositions* are justified; I have not addressed the question of whether *people* are justified in believing this or that proposition. My own mental state can be an indicator of the mental states of others, whether or not I know that this is true, or understand why it is true. But what is required for people to be justified in extrapolating from Self to Other? Must they know the fine points of phylogenetic parsimony or of likelihood analysis? Or does it suffice that a given extrapolation is sanctioned by parsimony or likelihood considerations? This is the epistemological thicket in which internalist and externalist views of justification are in contention.[35] I shall not try to evaluate these different approaches to the concept of justification, nor to say in any detail how they are related to the ideas I have developed in this paper, but I do want to explore one line of questioning that arises from an internalist point of view.

---

[35] For discussion, see Laurence BonJour, "Externalism/Internalism," in Jonathan Dancy and Ernest Sosa, eds., *A Companion to Epistemology* (Cambridge: Blackwell, 1992).

Is it possible for me to figure out that proposition (P) is true, if I do not already know whether Other has *M* or *A*? Surely, I can. I can tell whether *M* is heritable by looking at still other individuals who are known to have either *M* or *A* and see how they are related genealogically. But is it possible to determine that (P) is true without knowing anything about which individuals (other than one's self) have *M* and which have *A*? Without this information, how can I tell whether the traits are heritable?[36] Well, knowledge in some strong philosophical sense is probably unnecessary, but perhaps judgments about heritability require one to have reasonable opinions, however tentative, about which individuals have which traits. Even so, the solution to the problem of other minds that I have suggested would not be undermined. The incremental version of the problem, recall, asks whether knowledge of my own case should make a difference in the characteristics I attribute to others. It is not required that I conceive of myself as beginning with no knowledge at all concerning the internal states of others.

We learned from Good that there is no saying whether a black raven confirms the generalization that all ravens are black unless one is prepared to make substantive background assumptions. The mere observation that the object before you is a black raven is not enough. The same point, applied to the problem of other minds, is that the mere observation that Self and Other share *B* and that Self has *M* is not enough. Further assumptions are needed to say whether these observations confirm the hypothesis that Other has *M*. Recognizing this point in the ravens paradox does not lead inevitably to skepticism, and it should not have that effect in the case of Self-to-Other inference. The problem of other minds should not be shackled with the "methodological fiction" that Hempel imposed on the ravens paradox. When the fetters are broken, the problem of other minds turns into the problem of searching out common causes.

ELLIOTT SOBER

University of Wisconsin/Madison
London School of Economics and Political Science

---

[36] To be sure, it is possible to tell whether the behavior *B* is heritable (since *B* is observable), but the heritability of *B* is neither necessary nor sufficient for the heritability of *M* and *A*. I thank Branden Fitelson and Richard Lewontin for helping me clarify this point.

REVIEW ESSAYS

The Foundations of Causal Decision Theory. JAMES M. JOYCE. New York: Cambridge University Press, 1999. xii + 268 p. Cloth $54.95.*

James Joyce has offered an informed and clear account of how controversies concerning rational choice ought to be understood from the point of view of advocates of causal decision theory. You are unlikely to read a better apology for causal decision than Joyce's book.

This high praise should be tempered by the caveat that causal decision theory is an extremely controversial account of rational choice. In spite of the considerable persuasive skill and substantial learning displayed by Joyce, he does not succeed in rendering causal decision theory any less controversial than it was before. His book does, nonetheless, provide an excellent occasion to revisit some of the reasons why many of us continue to think that causal decision theory ought not to be taken seriously as a prescriptive theory of rational choice.

Like other causal decision theorists, Joyce sees the opposition to causal decision theory to be a view dubbed "evidential decision theory." This vision is extremely misleading. Evidential decision theory as understood by Joyce is the approach developed by Richard C. Jeffrey in his famous *The Logic of Decision.*[1] Even if attention is restricted to those who think that some form of expected-utility maximization lies at the core of rational choice, neither evidential nor causal decision theory get much of a press among statisticians, economists, management scientists, psychologists, and other nonphilosophers interested in rational choice. Authors like Frank P. Ramsey, Bruno De Finetti, Leonard J. Savage, Peter C. Fishburn, Francis Anscombe and Robert J. Aumann, Morris De Groot, Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld—to name but a few—endorse some form of expected-utility maximization but are neither evidential nor causal decision theorists.

Joyce, of course, knows this; but he thinks that Jeffrey's theory has replaced the more standard versions of expected-utility theory so that

---

* Thanks are due to Horacio Arló Costa, Philip Kitcher, Sidney Morgenbesser, and Teddy Seidenfeld for their helpful comments.

[1] New York: McGraw Hill, 1965.

0022-362X/00/9707/387–402

the serious confrontation is between Jeffrey's evidential decision theory and causal decision theory. It is to Joyce's credit in contrast to many other causal decision theorists that he seeks to argue for the replacement of more classical approaches by Jeffrey's proposal. He devotes considerable space to a discussion of the classical approach of Savage to securing a representation of a decision maker's preference among potential options in terms of expected utility.[2]

Once he has satisfied himself that Jeffrey's evidential approach has replaced Savage's, Joyce discusses the confrontation between evidential and causal decision theory. He offers improved representation theorems for both theories. In doing so, he changes the bases from which such representations are derived without seriously addressing the issue of why changing the bases constitutes improvement.

Finally, Joyce claims that causal decision theory is more sensitive than evidential decision theory to causal considerations in practical deliberation. Whether this is true or not, neither evidential decision theory nor causal decision theory displays the sensitivity to the question of what, from the decision maker's point of view, is subject to the decision maker's control which the view of Savage and other classical expected-utility theorists display. Insofar as we are sensitive to this kind of causal consideration, we should be more reluctant than Joyce is to dismiss classical views such as Savage's.

Some gesture in the direction of an elaboration of these judgments on Joyce's rich discussion is in order.

Joyce sees both causal and evidential decision theory as formal instantiations of *conditional expected-utility theory* according to which the value $V(X||A)$ of proposition X conditional on the supposition that the decision maker will perform proposition A is given by the following equation:

$$(1)\quad V(X||A) = \Sigma_S[P(S\&X||A)/P(X||A)]u(O[A,S])$$

In equation (1), $O[A,S]$ is a proposition that is a logical consequence of the decision maker's state of full belief K expanded by adding the proposition $A\&S$ to form $K^+_{A\&S}$.[3] A is the proposition that a given action that is optional for the decision maker (according to the information in K) is implemented. S is a member of a set W of propositions such that both K and $K^{+}{}_{A}$ entail that exactly one member of W is true and where each member is consistent with both K and $K^{+}{}_{A}$. X is a disjunction of elements of W. $O[A,S]$ describes everything of value to the decision maker which follows from $K^{+}{}_{A\&S}$.

---

[2] Savage, *The Foundations of Statistics* (New York: Wiley, 1954).

[3] The notion of a state of full belief is expressed in my terminology and not Joyce's. He examines states of probability judgment. As far as I can tell, a state of full belief as I conceive it corresponds to the set of propositions the agent "takes to be facts"—that is to say, judges to be probable to degree 1 according to a state of probability judgment. The fit is not perfect. In my view, fully believing that S is true is sufficient but not necessary for assigning S probability 1. So there is a difference between being absolutely certain (fully believing) that S and being almost certain which is not represented by states of probability judgment that use probability

Focusing on cases where $X = A$, equation (1) becomes the following:

$$(2)\quad V(A||A) = \Sigma_S P(S||A)\,u(O[A,S])$$

When $P(S||A)$ is replaced by the conditional probability operator $P(S/A)$ equal to $P(S\&A)/P(A)$ when $P(A) > 0$ but well defined for all A consistent with K even when $P(A) = 0$, (2) becomes the following:

$$(3)\quad V(A) = \Sigma_S P(S/A)\,u(O[A,S])$$

The conditional probability $P(x/y)$ relative to K has additional significance above and beyond that afforded by the partial definition just given. Consider any expansion $K^{+}{}_{y}$ of K by adding y consistent with x. The unconditional probability $P^{+}{}_{y}(x)$ relative to $K^{+}{}_{x}$ is determined by K, x, and P relative to K according to the requirement that $P^{+}{}_{y}(x) = P(x/y)$. In the context where K is the agent's state of full belief and P represents his state of subjective-probability judgment, $P^{+}{}_{y}$ can be thought of as the agent's state of subjective-probability judgment on the supposition that y is true. In particular, the decision maker that has not as yet made up his mind supposes purely for the sake of the argument that option A is implemented and considers the probability that S on that supposition. The decision maker expands K by adding A (consistent with K) to A and forming the deductive closure. The probability is then adjusted according to the recipe just indicated.[4]

---

functions from propositions to values in the standard real numbers. Joyce does not appear to consider "almost certainty" in his discussion. With this qualification, the characterization of supposition given in the text captures informally and simply the ideas concerning supposition addressed by Joyce using the technically more complex ideas of section 6.3 of his book. No substantial issue being discussed here is begged by this modification of Joyce's presentation.

[4] I focus here on supposition when the deliberating agent is in suspense as to whether the supposition is true or false. This is the main context in which a deliberating agent supposes that an option is chosen and implemented. This account of supposition agrees with Frank Ramsey's account of conditionals for the open case. If $K^{+}{}_{A}$ entails S, the Ramsey Test as Ramsey formulated it dictates that the "conditional" $A > S$ is acceptable at K and conversely. When $K^{+}{}_{A}$ does not entail S, S has a probability relative to $K^{+}{}_{A}$ equal to $P(S/A)$ as explained in the text. $P(S/A)$ could equal 1 even though $\neg S$ is consistent with $K^{+}{}_{A}$.

Evidential decision theorists maximize expected utility according to formula (3). According to Joyce, in so doing, the decision maker is maximizing the *auspiciousness* of her choice. Choice is based on which option confirms the best kind of news.

Maximizing expected utility according to (3) distinguishes evidential decision theorists from causal decision theorists. It is important to emphasize (as Joyce does not) that all decision theorists in the classical school would endorse (3) as long as decision problems can be formulated in state functional form where outcomes are uniquely determined by the implementation of options and the truth of states. I shall return shortly to the difference between evidential and classical decision theory.

Causal decision theorists proceed differently. For them, to suppose that A is the case relative to belief state K is to proceed as follows: for every possible world w doxastically possible according to K, consider the set $\#A(w)$ of worlds nearest or most similar to w at which A is true and divide the total probability $P(w)$ equally among each of these to obtain $P\#A(w)[w^*]$ for w^* a member of $\#A(w)$. This then yields a probability $P\#A(w)[S]$ where the proposition that S is represented by the set of worlds at which S is true. This probability will be the sum of the values for $P\#A(w)[w^*]$ for w^* a member of both $\#A(w)$ and S. $P(S\backslash A) = \Sigma_w P\#A(w)[S]$.

When $P(S||A)$ is replaced in (2) by $P(S\backslash A)$, we obtain a result apparently different from (3).

$$(4)\quad U(A) = V(A\backslash A) = \Sigma_S P(S\backslash A)u(O[A,S])$$

According to Joyce, the expected value given in (4) measures the efficaciousness of performing A rather than its auspiciousness. This is asserted without any argument, as far as I can tell.

Joyce interprets both $P(S/A)$ and $P(S\backslash A)$ to mean the subjective probability that S is true on the supposition that A is true. He contends that there is an ambiguity in the notion of supposition. The difference derives from a contrast in the way the initial state of full belief K is transformed when A is added by supposition to K. Probability judgments conditional on the supposition are derived from the probability judgments relative to K according to the transformations previously sketched. Prior to reaching a decision, the deliberating agent is in suspense as to whether A is true or false so that A is consistent with K. $K^+{}_A$ is consistent and the unconditional probability that S relative to $K^+{}_A$ is taken to be equal to $P(S/A)$ relative to K. This is the kind of supposition Joyce calls *indicative supposition.* By way of contrast, *subjunctive supposition* involves taking the image $K\#A$ of K

by adding A to K and grounding probability judgment relative to the image. This yields $P(S \backslash A)$.

How different is supposing that A via imaging from supposing that A via AGM or Ramsey in the open case (the only case to be considered here)? $K\#A$ can in that case differ from $K^{+}{}_{A}$ quite markedly. Indeed, it is in the open case that the structural difference between imaging and AGM revision is the clearest. Nonetheless, imaging can give the same verdict as expansion if nearest A-worlds to doxastically possible $\neg A$-worlds (relative to K) are restricted to doxastically possible worlds. When this happens, it appears that $P(S \backslash A) = P(S/A)$. In that case, causal decision theory should deliver the same verdict as evidential decision theory.

This is fortunate because even the most ardent causal decision theorists are inclined to acknowledge that in a wide variety (and, perhaps, almost all) contexts, imaging and expansion give the same verdicts in decision making. Causal decision theorists invoke this circumstance to explain why it took Robert Nozick's[5] discussion of Newcomb's problem to bring to consciousness that there is a serious difference between causal decision theory and evidential decision theory. In Newcomb's problem, in the gene for lung cancer and smoking problem and the twin prisoners' dilemma, $P(S/A) \neq P(S/B)$ for at least one available option B. I focus here on supposition when the deliberating agent is in suspense as to whether the supposition is true or false. This is the main context in which a deliberating agent supposes that an option is chosen and implemented.

Suppositional reasoning in the open case agrees with Ramsey's account of conditionals for the open case. If $K^{+}{}_{A}$ entails S, the Ramsey Test as Ramsey formulated it dictates that the "conditional" $A > S$ is acceptable at K and conversely. When $K^{+}{}_{A}$ does not entail S, S has a probability relative to $K^{+}{}_{A}$ equal to $P(S/A)$. Yet, according to causal decision theorists, S is causally independent of A in the sense that $P(S \backslash A) = P(S \backslash B)$ for every alternative available option B. Consequently, $P(S \backslash A) \neq P(S/A)$ for at least one available option A. Hence, supposition by imaging is not equivalent to supposition by expansion.

If Newcomb-type problems do not provide a basis for distinguishing supposition by imaging from supposition by expansion in the open case, the contrast disappears. This accounts for the importance of the Newcomb-type cases for causal decision theorists and for philosophers who insist that supposition via imaging is more than a metaphysician's plaything.

---

[5] "Newcomb's Problem and Two Principles of Choice," in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel* (Boston: Reidel, 1969), pp. 114-46,

Appeal to intuitions about Newcomb's problem and the prisoner's dilemma for twins is a very weak reed on which to rest the case for causal decision theory. Yet it seems to be the only serious independent source to which causal decision theorists can appeal for support of their doctrine. Joyce unabashedly makes the appeal without considering the reservations of critics.

Consider the prisoners' dilemma for twins. Each prisoner judges that it is highly probable that she and her twin will both choose to confess or both choose not to confess. That is part of what it is to judge that one is very much like one's twin. Causal decision theorists read this erroneously as implying that from prisoner 1's point of view, the conditional probability of prisoner 2's confessing (not confessing) given that prisoner 1 confesses (does not confess) is very high. That is to say, given that $P(C_1\&C_2) + P(\neg C_1\&\neg C_2)$ is very close to 1, it is concluded that $P(C_2/C_1)$ and $P(\neg C_2/\neg C_1)$ are both near 1. Maximizing expected utility according to (3) then recommends refusing to confess. Maximizing expected utility according to (4) recommends confessing on the grounds that prisoner 1 judges that what prisoner 2 will do is causally independent of what prisoner 1 does.

A quick calculation, however, reveals that concluding that $P(C_2/C_1)$ and $P(\neg C_2/\neg C_1)$ are both near 1 is fallacious. Consider a case where $P(C_1\&C_2) \approx 1$. $P(\neg C_2/\neg C_1)$ could very well be close to 0. In that event, maximizing expected utility according to formula (3) *recommends confessing*. This option is, of course, the dominating option favored by causal decision theory. To be sure, there are some cases where using formula (3) favors the dominated option. The prisoner's dilemma for like-minded twins does not specify whether the conditions relevant to discriminating between cases where (3) favors confessing and (3) favors not confessing are in force. Indeed, someone confronting this kind of puzzle might think of it as a situation where the probabilities relevant for computing expectations are indeterminate. The prisoners are then facing a decision problem under uncertainty where everyone agrees that the dominating option, confessing, is to be recommended. The interpretation favored by Joyce, according to which there should be a conflict between the recommendations of (3) and (4), is not implied by the problem when used as an intuition pump any more than are any of the alternative interpretations to which I have gestured.

To adopt an interpretation that yields such a conflict presupposes that the unconditional probability of either prisoner confessing is approximately 0.5. Can anyone honestly say that presystematic judgment clearly comes down in favor of confessing whether or not this is so? Unless causal decision theorists can make a compelling non-question-begging case for saying so in this and other Newcomb-type cases, the rationale for pursuing the foundations of causal decision theory along the lines developed by Joyce just is not there. He seems serenely confident that presystematic judgment is on his side. Serene assertion is no substitute for persuasive argument.[6]

Let serene assertion carry the day anyhow. There are more serious difficulties. Let us allow Joyce and other causal decision theorists directly to assume that $P(C_2/C_1)$ and $P(\neg C_2/\neg C_1)$ are both near 1. But if prisoner 1 adopts these conditional probabilities because prisoner 2 is his twin, presumably prisoner 1 should also judge that $P(C_1/C_2)$ and $P(\neg C_1/\neg C_2)$ are both near 1 because prisoner 1 is the twin of prisoner 2.

Under these circumstances, the unconditional probability $P(C_2)$ prisoner 1 assigns to prisoner 2 confessing is (as just noted) near 0.5. This is a consequence of the calculus of probabilities. Another consequence is that $P(C_1)$ is near 0.5.

The numbers do not matter. What is important is that prisoner 1 is committed, by his conviction that he is one of a pair of twins in a prisoner's dilemma, to assigning unconditional probabilities to whether he confesses or not. This commitment is derived from the requirement that $P(C_1/C_2)$ and $P(\neg C_2/\neg C_2)$ both are near 1 as are $P(C_2/C_1)$ and $P(\neg C_2/\neg C_1)$. All four quantities are assumed to be well defined. The calculus of probabilities then requires that $P(C_1)$ and $P(C_2)$ both be well defined. Hence, prisoner 1 is committed to assigning unconditional probabilities to her own acts.

In the Newcomb problem, the decision maker is also committed to assigning unconditional probabilities to the available options. The same is true in every Newcomb-type problem.

Now, according to Joyce, the rival to causal decision theory is Jeffrey's evidential decision theory. According to Jeffrey's theory, every proposition in the algebra and, hence, every proposition used to represent option propositions, outcome propositions, and state propositions is assigned an unconditional probability.

So causal and evidential decision theorists agree that the decision maker is committed to assigning unconditional probabilities to the propositions representing available options. By way of contrast, it is not an ambition of Savage's approach to rational choice to assign unconditional probabilities to options. Nor is it an ambition of most other authors engaged in laying down postulates for deriving probabilities and utilities from preference among potential options.

---

[6] This is the gist of the view of Newcomb's problem I took in my "Newcomb's Many Problems," *Theory and Decision*, VI (1975): 161-75.

Thus, no matter what views they may have held concerning the issue, authors who are neither causal decision theorists nor evidential decision theorists have embraced accounts of expected utility that do not commit them to assigning unconditional probabilities to options.

Moreover, several authors (Frederic Schick, Wolfgang Spohn, Kadane, and Teddy Seidenfeld, and myself[7]) have argued that deliberation crowds out prediction so that a decision maker may not coherently assign unconditional probabilities to the propositions he judges optional for him. Savage et alia do not mandate this view explicitly. They do not preclude it. Advocates of evidential and causal decision theories do. This leads to trouble.

Principles of rational choice are supposed to be nonvacuously applicable by the decision maker to the evaluation of his options at a point where the agent has become clear as to what the relevant information is concerning his options, probability judgments, and utility judgments. If the decision maker is certain at that stage that he will choose an optimal option (that is to say, choose rationally), the only options available to him must then be optimal. If choosing that proposition A be true is optional for agent X, X should be convinced that, if X chooses A, A will be true. Hence, if X is convinced that X will choose optimally, X is convinced that every proposition describing a suboptimal course of action will be false. Suboptimal courses of action will have been ruled out as possibilities and, hence, as available options for choice. There will be no circumstance under which the deliberating agent X can apply the principles of rational choice to his predicament in order to reduce a large set of available options to a subset of optimal ones.

Matters are not improved if one allows the decision maker to assign probabilities to the option propositions. If the probabilities assigned have any relevance to decision making, decision makers should be absolutely certain that they will choose optimally.[8] They should rule out suboptimal propositions as optional. Once more, criteria for rational choice become vacuous.[9]

---

[7] Schick, "Self Knowledge, Uncertainty and Choice," *British Journal for the Philosophy of Science*, xxx (1979): 235-52; Spohn, "Where Luce and Krantz Do Really Generalize Savage's Decision Model," *Erkenntnis*, xi (1977): 113-34; Kadane and Seidenfeld, "Equilibrium, Common Knowledge and Optimal Sequential Decisions," reprinted as chapter 32 of *Rethinking the Foundations of Statistics*, Kadane, Schervish, and Seidenfeld, eds. (New York: Cambridge, 1999); and my *Hard Choices* (New York: Cambridge, 1986), chapter 4; and *The Covenant of Reason* (New York: Cambridge, 1997), chapters 2, 4, and 5.

To avoid this unfortunate result, we should conclude that although the deliberating agent may predict his future choices as well as the decisions of others, a self-critical agent cannot coherently assign unconditional probabilities to his currently available options.

For authors embracing this view, the evidential decision theory of Jeffrey and causal decision theory must both be unacceptable as accounts of prescriptive principles of rational choice applicable by self-critical deliberating agents to the policing of their own deliberations. Evidential and causal decision theories are insensitive to the distinction between what is under the decision maker's control and what is not.

---

[8] Let option *A* be strictly preferred to option *B*. If the decision maker can coherently assigned probabilities to *A* and *B*, he should be in a position to accept (or reject) bets on *A* while choosing *A* true and while choosing *B* true. Since the decision maker judges choosing *A* (*B*) to be under her control, the decision maker should be convinced that relative to her state of full belief, choosing *A* (*B*) true implies *A* (*B*). Given that *A* and *B* are the only two options prior to being offered the bets, the decision maker should be prepared to pay any price up to the stake *S* for a bet on *A*. According to the decision maker, the fair betting rate for the gamble should be 1. The subjective probability for the decision maker's choosing *A* should be 1, and this should stand for absolute certainty. If there are several optimal options in a decision problem, the decision maker should be absolutely certain that she will choose one of them. This argument could fail if the probability assigned to an option is not connected with the assessment of the expected utility of choosing that option. To detach subjective probability from its role in determining expected value would, however, seem a far more radical move than any antibehaviorist might be prepared to endorse. See footnote 9.

[9] Jeffrey—"A Note on the Kinematics of Preference," reprinted as chapter 16 in *Probability and the Art of Judgment* (New York: Cambridge, 1992)—worries as to whether the deliberating agent must be certain that he will choose an optimal option. He resolves the problem by distinguishing rightly between the agent's views before "the time for action" and upon taking action. Joyce endorses this view. The issue of concern is, however, whether before action is taken, the decision maker is in a position to determine which of the options available to him, as he judges at the time, are optimal for him according to the values he endorses at that time. I argue as follows. Suppose the agent is coherent in his probability and utility judgments and is certain that he will choose among the propositions he judges to be feasible for him on the basis of the beliefs and preferences he has prior to taking action. Suppose further that unconditional probabilities are to be assigned to his options. On these suppositions, he should be certain that he would choose optimally. Jeffrey does not deny this. He suggests rather that either the decision maker does not as yet recognize what his options are or is not sure that he will not change his preferences. But this means that Jeffrey admits that there is no context in which the decision maker can apply the injunction to maximize expected utility among the available options to reduce the set of available options to a proper subset of optimal options. That is to concede that the principles of rational choice cannot be applied nonvacuously to reduce a feasible set to an optimal subset. The concession is death to decision theory as a set of principles to be used by self-critical agents in practical deliberation. The result coheres well with a "Logic of News Value" but not with a "Logic of Decision."

If this is right, then none of the scenarios where causal decision theory is alleged to yield a verdict according to (4) different from expected-utility theory according to (3) is a predicament where the agent has a choice between alternative options. Neither of the twin prisoners has any options. Confessing or failing to confess is not under the prisoner's control, according to the prisoner's own view of her predicament. The alleged difference between $P(X\backslash Y)$ and $P(X/Y)$ has no application!

And the main rival to causal decision theory is not evidential decision theory. Evidential decision theory suffers from the same insensitivity to causal structure as causal decision theory does. The classical account of expected utility articulated in difference ways by Ramsey, De Finetti, John von Neumann and Oskar Morgenstern, Kenneth Arrow, Savage, Fishburn, Anscombe and Aumann, among many others, need not prohibit the assignment of unconditional probabilities to options. But this account is compatible with such a prohibition. Spohn, Schick, Kadane and Seidenfeld, and I have walked the extra mile and suggested that assigning unconditional probabilities to options is not rationally coherent.

Before concluding that Joyce's vision of a confrontation between evidential and causal decision theory is without justification, Joyce's criticisms of Savage should be examined.

Consider a decision maker facing a choice between options A_1, $A_2,\ldots,A_n$. The decision maker wishes to evaluate A_i in terms of (a) the values or utilities assigned to possible outcomes o_{i1}, $o_{im(i)}$ (each o_{ij} is consistent with the expansion of K by adding A_i and that expansion entails that exactly one of the o_{ij}'s is true); and (b) the probability distribution $P(o_{ij}/A_I)$. According to all variant forms of expected utility-theory except causal decision theory, the value of A_i should be equal to the following:

$$(5)\quad V(A_i/A_i) = \Sigma P(o_{ij}/A_i)u(o_{ij})$$

The virtue of using (5) is that it invokes consideration of the options available to the decision maker and possible outcomes of these options, the conditional probabilities of the outcomes on the supposition that the options are true, and the utilities of the outcomes.

There is no need, in general, to invoke a partition of the space of doxastic possibility into states and to require that the partition be such that o_{ij} is uniquely determined, given K, by A_i and a state S_j according to a function $O[A_i,S_j]$. Nor is there a need to assign unconditional probabilities to either the option propositions or the outcome propositions.

Expected utility as characterized by (5) is often regarded as the most general version of expected-utility theory. It was formulated as such by Fishburn.[10] Unfortunately, no one knows how to derive the utilities for outcomes or conditional probabilities requisite for applying (5) from information about preferences about the available options or an extension of such preferences to a domain of potential preferences. Measuring probability and utility on the basis of information about preferences among potential options calls for invoking additional assumptions about the structure of the decision problem. In general, a partition of the space of doxastic possibility into states is supplied such that each $o_{ij} = O[A_i,S_j]$. Given such a partition, the options are representable as functions from states to outcomes and the decision problem is representable in *state functional form.* Given this, equation (5) reduces to equation (3) of so-called evidential decision theory that Joyce equates with Jeffrey's theory. The conditional probability $P(o_{ij}/A_i)$ becomes $P(S_j/A_i)$. In contrast to Jeffrey's theory, however, there is no appeal to unconditional probabilities for the option propositions like A_i. Neither the general principle (5) formulated by Fishburn nor the formula (3) entails a commitment to evidential decision theory.

Savage's approach to deriving expected-utility representations of preferences among available options from preferences among potential options (Savage acts) is based on three assumptions whose joint satisfiability in all applications is very much open to question.

(1) The options actually available in realistic decision problems can be represented as functions from states in a set W to outcomes in a set O. That is to say, they are representable in *state functional form.*
(2) The set of *potential* options (acts), over which a preference ranking is given, consists of all propositions representable by functions from W to O. The preference ranking over the options actually available is embedded in this preference ranking.
(3) The preference ranking over the acts or potential options satisfies Savage's postulates.

Joyce misconstrues Savage as presupposing the availability as actual options of all acts or potential options representable as functions from the given stock of states to the given stock of outcomes. The presupposition would entail recognizing a gamble as an available option where the payoff for heads is eternal peace and the payoff for tails is thermonuclear war. Like Jeffrey, Joyce rightly takes this to be absurd. Joyce finds "constant" acts especially incomprehensible.

---

[10] *Decision and Value Theory* (New York: Wiley, 1964).

Joyce (like Jeffrey) is, indeed, misconstruing Ramsey, Savage, and other classical decision theorists. Savage's approach does not require that the preference ranking over potential options be a preference ranking over actual options any more than theories of consumer demand based on indifference maps over commodity bundles require us to think of all the commodity bundles as actually available to the consumer. There is textual evidence that Savage[11] clearly understood this. In efforts to elicit an agent's probabilities and utilities, the decision maker is invited to consider how he would make choices in various hypothetical decision problems where the options are potential options generated from the options the decision maker actually faces in the manner specified by Savage.

Savage, like Joyce, is fully prepared to engage in suppositional reasoning even though the kind of reasoning he allows is quite different from Joyce's. The charge Joyce levels against Savage—that his consideration of a preference ranking including constant acts involves an egregious lack of realism—looks plausible if Savage is interpreted as claiming that constant acts are always available options in realistic decision problems. But this interpretation fails to do justice to Savage's own intentions or to the more general problem of developing a theory of elicitation eventuating in representations of the preferences over potential acts by unconditional probability assignments to states and utility assignments to outcomes.

Joyce's complaint misses its mark. Even so, the "realism" of Savage's approach can be questioned. Savage's approach as sketched in conditions (1), (2), and (3) does imply that not only are the options actually available representable in state-functional form but that the states ingredient in that representation are probabilistically independent of the available options. It is far from obvious that in realistic applications, decision problems can always be represented in this manner. Fishburn's suggestion that states be expressed by conditionals presupposes the dubious assumption that conditionals carry truth values. And it does not secure probabilistic independence.

Nonetheless, in those contexts where state-functional form representability with probabilistic independence can be secured, Savage's formalization promises an account of how preferences among potential options can be used to derive probability-utility pairs for representing preferences among options in terms of expected utility. Joyce himself endorses the widely held view that preferences meeting Savage's requirements guarantee that the probability components in the set of probability-utility pairs ensuring expected-utility representation (1) are one and the same and that the utility components are unique up to a positive affine transformation.

---

[11] See the letter of January 27, 1987 from Savage to Robert Aumann, reprinted by Jacques Drèze in Appendix A of chapter 2 of *Essays on Economic Decisions under Uncertainty* (New York: Cambridge, 1987), especially the top two paragraphs on p. 80.

The claim about the uniqueness of the representation is not accurate. There are nondenumerably many other probability-utility pairs that will represent the same preference over the acts and satisfy Savage's requirements. The utility of an outcome can change, however, depending upon which state (or event) determines the outcome for some act or other. Nothing in Savage's axioms for preferences over acts precludes this or can preclude this.[12]

Joyce fails to appreciate this point. Since Savage and others had thought of the outcome o_{ij} presystematically as describing all those features of the world which the decision maker cared about, Joyce takes it to be faithful to Savage's project to stipulate that the same proposition cannot describe the outcome of option A_i in state S_j and of option A_{i*} in state S_{j*} unless the two outcomes are equipreferred. Savage's project, however, prohibited making such assumptions about equipreference between outcomes except insofar as they could be derived from preference among potential options. Interestingly enough, Savage's approach cannot make the required derivation.

The problem of state-dependent utility does constitute a genuine difficulty for Savage's project. It may be granted that Savage's theory cannot yield a technique for measurement of probabilities and utilities from preferences over potential options which covers every kind of decision problem. Still, as long as it is conceded that the theory of measurement is applicable in a nontrivial class of cases, its importance will have been established. The issue of state-dependent utility calls into question whether Savage's theory of measurement is applicable in a nontrivial class of cases.

---

[12] Ramsey, who undoubtedly wanted his account to be one where the basic notion of preference was revealed by choice among uncertain prospects, explicitly acknowledged that his introduction of ethically neutral propositions was intended to deal with a difficulty confronting his project. The ethical neutrality of a proposition h is not derived from preferences among potential options but by considering any proposition g such that both h and $\neg h$ are consistent with g and requiring that gh and $g\neg h$ be equipreferred. It seems that Ramsey thought that once he defined probability ½ in an ethically neutral proposition, he would have no further worries about state-dependent utility. On this score, he was mistaken—Ramsey, "Truth and Probability," in *Philosophical Papers*, D.H. Mellor, ed. (New York: Cambridge, 1990), pp. 52-101, see especially pp. 72-74. A propos the topic of state-dependent utility, in the same letter to Aumann referenced in footnote 11, Savage acknowledges the tie between his worry about small versus large worlds and the question of state-dependent utility. Joyce misses this connection in his discussion of Savage on this topic (section 3.4).

Thus, Joyce's critique of Savage is misdirected simply because Joyce does not take seriously the project for which Savage's theory was designed—to wit, to derive unconditional probabilities for states and utilities for outcomes from preferences over potential options (that is, Savage acts). Nonetheless, it is apparent that Savage has not succeeded in realizing his own project. Joyce rightly adopts this view, but for the wrong reasons.

Joyce sees Jeffrey-Bolker theory as a significant advance over Savage. What is that advance? It cannot be that Jeffrey has succeeded in deriving probabilities for states and utilities for outcomes from preferences over potential options. The representation theorem offered derives probabilities for all propositions in the given algebra and utilities for all propositions with the exception of the inconsistent proposition from preferences over the entire algebra minus the inconsistent proposition. It seems farfetched to think of the space of potential options as being the entire algebra of propositions characterizing the conceptual space. Jeffrey has not made progress in realizing the project undertaken by Savage. If an advance has been made, it is by turning to a better research project. What is better about Jeffrey's research project?

If Joyce's concern is with a decision theory exhibiting improved sensitivity to causal considerations, Savage's theory is clearly superior to Jeffrey's. A distinction between what is under the decision maker's control and what is not is built into the structure of the problem of representation which is absent from Jeffrey's account.

Jeffrey's project employs "data" concerning the agent's preferences for "news" in lieu of data concerning preferences for potential options to derive judgments concerning credal probability and utility. This looks like a backward step if we have interest in sensitivity to causal structure. And it does not help us. Given maximally specific and determinate information about preferences over propositions in the algebra, we still cannot derive a unique probability and a utility unique up to a positive affine transformation.

Joyce recognizes a weakness in Jeffrey's program on this score. His solution is to abandon the project of deriving probability-utility representations from preferences over propositions *alone*. He proposes adding to the database information about comparative probability judgments satisfying additional postulates. Does this supplement render Jeffrey's theory superior to Savage's? Schervish, Seidenfeld, and Kadane observe that one can use DeGroot's theory of comparative probability to supplement Savage's theory in a way that identifies a unique probability distribution for states and settles the question of state dependent utility. They also take note of the approach proposed by E. Karni, D. Schmeidler, and K. Vind and show that it, too, yields a unique probability distribution for states.[13] There is no advantage in Jeffrey's theory to be found here.

We have new questions to ponder. Why all the fuss about finding representation theorems? I do not mean to question the valuable structural and mathematical insights to be gained by exploring the possibilities of representation theorems with different bases. These are not to be dismissed. But Joyce insists that the construction of a representation theorem is crucial to an adequate decision theory no matter what the basis is from which it is derived. He declares his main project is to provide a representation theorem for causal decision theory. Joyce's improved representation theorems for evidential and causal decision theories are the *pieces de resistance* of his interesting book. But the representation theorems he considers are a heterogeneous lot derived from different types of bases. What warrants comparing them for better or worse as Joyce does? Why should we not rest content with specifying the structures of decision problems as best we can, indicating where probability and utility representations are needed in order to obtain expected utilities for options? We can do this without deriving the numerical probability and utility representations from preferences (either for potential options or over the entire domain of propositions) and without supplementation by judgments of comparative probability. Once we have given up the concern to derive probability and utility functions from potential data about choice, why will this attitude not do? Joyce does not give a clear answer.

In the absence of an answer, we might be tempted to claim that Fishburn's early general formulation of the expected-utility principle modified to accommodate indeterminate probabilities and utilities is an appropriate "foundation" for decision theory.

Succumbing to such a temptation does call into question the relevance of imaging-based supposition to rational deliberation. But doing so does not argue in favor of evidential decision theory as opposed to causal decision theory. Practical wisdom calls for a proper understanding of the difference between what is under our control and what is not. Causal decision theorists are right to complain about the insensitivity of evidential decision theory to causal considerations. Causal decision theorists are, however, no more sensitive to the relevant aspects of causal structure than evidential decision theorists are. Neither evidential decision theorists nor causal decision theorists appreciate that deliberation crowds out prediction. A sensible account of practical deliberation should look beyond the debate between causal and evidential decision theorists.

---

[13] Schervish, Seidenfeld, and Kadane, "State Dependent Utilities," *Journal of the American Statistical Society*, LXXXV (1990): 840-47; DeGroot, *Optimal Statistical Decisions* (New York: Wiley, 1970); Karni, Schmeidler, and Vind, "On State Dependent Preferences and Subjective Probabilities," *Econometrica*, LI (1983): 1021-31.

ISAAC LEVI

Columbia University

BOOK REVIEWS

The Dappled World: A Study of the Boundaries of Science. NANCY CARTWRIGHT. 247 p. New York: Cambridge University Press, 1999. Cloth $54.95, paper $19.95

This is a collection of nine essays with an extended introduction. The essays primarily treat laws, models, and theories with applications drawn mostly from physics and economics. Some essays are accompanied by lovely illustrations (Rachel Hacking gets first prize here) and some of the essays are new or contain new material. The approach, familiar from Nancy Cartwright's previous publications, is empiricism laced with natural capacities. This leads her to view science as an untidy patchwork of laws for a world with features that are fully ordered, roughly ordered, or not ordered at all—the dappled world of the title.

Contrasting with a dappled world is one that is universally ordered. The world of physics portrayed by a grand unified theory would be one such case, the world of bio-medicine portrayed by the genetic paradigm, another. These constitute a "fundamentalism" that Cartwright opposes. In framing that opposition, she employs ordinary-language philosophy, interestingly bolstered by G. E. Lessing's analysis of Aesop's fables. On that basis, she draws a distinction between abstract concepts (such as quantum Hamiltonians, classical force functions, or electromagnetic field vectors) whose relation to the world is mediated by more concrete ones, and concrete concepts whose specific forms are provided by interpretive physical models. Thus, in $F = ma$, she regards force as abstract, needing the content provided in models, such as the linear harmonic oscillator. The upshot is that the concept of force has content (and the force law is true) only in the context of certain models. Fundamentalism, then, becomes a matter of faith that such models cover the whole world, and also faith that in all such models, the objects have a common property of being-subject-to-a-force. Cartwright believes fundamentalists need to posit such new properties in order to make all the instances that fall under an abstract concept (like force) resemble one another. This is the metaphysics of conceptual realism, which she would rather avoid.

Cartwright prefers another metaphysics. To do justice to the fact that laws are ceteris paribus, she thinks we need an account metaphysically richer than that provided by Humean regularities. Developing the line begun in her previous writings, she thinks we need natures or capacities.

In the Humean account, "a law of nature is a necessary regular association between properties antecedently regarded as OK" (49). The problem with Humean laws is that, by and large, they are false. If, to take one of Cartwright's examples, Coulomb's law (as a Newtonian force law) is understood as a Humean regularity that specifies the accelerations of each pair of charges, then the law is false, since the acceleration of most charges is not even close to that postulated by Coulomb's law. To preserve the (at least approximate) truth of Coulomb's law, we have to understand the law as containing an implicit ceteris paribus clause: if the masses of the charges are negligible and there are no other disturbing influences, then the two charges accelerate according to Coulomb's law. Cartwright expresses her thesis that laws hold only ceteris paribus in terms of what she calls *nomological machines*. A nomological machine (for instance, a battery) is a shielded arrangement of components that gives rise to law-like regularities. But it is difficult to see what real work this notion can do, since her thesis that laws hold only relative to the operation of a nomological machine appears to reduce to the claim that laws hold where they do.

However expressed, Cartwright focuses on two difficulties for the Humean. First, since a Humean can refer only to measurable or occurrent properties (the OK ones), "it is impossible [on that account] to find any appropriate category that would single out conditioning factors for laws from the factors that fall within them" (137-38). Second, even if a Humean account were to be able to accommodate ceteris paribus laws, it is faced with the problem that such laws hardly ever apply, since the conditions specified in the ceteris paribus clause are rarely satisfied. Read as mere regularities between occurrent properties, ceteris paribus laws tell us very little about the world.

Cartwright appeals to capacities to resolve both problems. First, she argues, the conditions specified in a ceteris paribus clause are distinct from the conditions governing the regularity. The former concern capacities, while the latter are OK, Humean properties. Second, knowledge of capacities allows us to extend our theories beyond the ideal circumstances in which the ceteris paribus conditions are satisfied and the Humean regularity holds. Thus, Cartwright concludes, "knowledge of capacities is more basic in that it is both more embracing and more widely useful than knowledge of regularities" (77).

Cartwright's arguments against the Humean are not convincing, however. First, the Humean could separate the conditioning factors for laws from the factors that fall within them roughly as follows. The conditioning factors are those which need to be held constant in order for the functional dependency to hold between the factors that vary. This way of drawing the divide need invoke no factors that are not OK. Second, to explain how we extend our scientific knowledge beyond the ideal circumstances in which the Humean regularities hold, the Humean might hold that complex systems are presumably governed by complicated laws and regularities, which science aims to discover, and that a strategy of using simple regularities as guides in the search for more complex ones is pragmatically justified.

Cartwright rejects this response as fundamentalist, as being at odds with her own picture of a dappled world. She argues that, since we do not have the complex, unifying laws that the Humean needs to invoke, the Humean's belief in them cannot be empirically founded. It is no more than a matter of faith and, thus, should have no place in an empiricist philosophy. But are good empiricists really locked into nothing more than local model building in a dappled world? The historical evidence suggests that at times science does succeed with highly local models, while at other times science advances through attempts to unify domains that were previously thought to be distinct. Empiricism, it seems, must allow for both.

Moreover, does Cartwright's appeal to capacities fare better than the Humean in explaining how it is that our scientific knowledge extends beyond the ideal circumstances in which the Humean regularities hold? Cartwright's natures or capacities differ from ordinary dispositions, such as fragility. In her conception, capacities are always *trying* to be realized. A glass does not try to break while it is not struck, but charged bodies always try to move according to Coulomb's law. Unfortunately, the notion of trying is rather vague, and in spelling it out, Cartwright sketches what appear to be two different and conflicting accounts. One account emphasizes the plasticity of capacities: while capacities are always trying to manifest themselves in some way or other, how they in fact manifest themselves can differ radically from one context to another. Cartwright cites the Hotelling economy (60ff.), where a single capacity associated with taxes can, depending on the circumstances, lead to either an increase or decrease in prices, or can leave prices unchanged. But if capacities are so open-ended, then it is not clear why an appeal to capacities and their manifestations in simple and tidy circumstances can help us anticipate what happens in other, untidy circumstances. Cartwright's answer is that knowledge of capacities suggests strategies for finding more complex regularities when conditions are less than ideal (59). Of course, this is just the Humean line dressed in capacities.

Sometimes, however, Cartwright suggests a different account of trying. Using "Coulomb's capacity" as a paradigm, she says that the capacity is "recorded in its own special law" and that "we know some very explicit rules for how Coulomb's capacity will combine with others" (54). Indeed, Newton's second law tells us how to combine the effects of additional forces on the charge to determine the resulting acceleration. Cartwright is clearly right in asserting that where they exist, such rules enable us to extend our knowledge beyond the simple circumstances encoded in our ceteris paribus laws. But this would seem to be a move open to the Humean as well. Arguably, rules that tell us how various capacities combine can be recast as involving only OK properties. In the case of Coulomb's law, Cartwright has an argument to block this move. The obvious candidate here for a Humean rule is Newton's second law, which relates acceleration to the combination of the component forces present. But as we have seen, Cartwright believes that force, being abstract, is not an OK property. For her, Newton's law is thus not a Humean law. Even if her argument for the abstractness of force is convincing, however, and even if abstractness makes force not OK, it is not clear whether the argument generalizes to block as not OK every rule for combining capacities. Finally, notice that the existence of general rules like Newton's law seems at odds with Cartwright's claim that the world is dappled. Thus, there appears to be a tension between two of Cartwright's main theses: that knowledge of capacities allows us to extend our scientific knowledge and that our world is indeed untidy and dappled.

We turn now from laws to models. Cartwright contrasts her view of the relation of models to theory with what she calls the *vending-machine view*. According to that view, all of the necessary resources for representing the world can be drawn more or less deductively from fundamental theory, so extracting representations of phenomena from theories is as easy as extracting candy from a vending machine. In a dappled world, however, theories do not straightforwardly represent local states of affairs. Rather, we need two kinds of models, applied in two successive steps: interpretive models and representative models.

Representative models are the final product. These models represent what actually happens in the world. To get from theory to representative model, however, we need to make use of an interpretive model, which, as we have seen, is the vehicle for applying theories that involve abstract concepts. According to Cartwright, theories come with a stock of interpretive models. (In the case of quantum mechanics, Cartwright recognizes just five—and physicists might worry about that.) So the first step in concrete applications of a theory is to choose from the stock a model that fixes an interpretation of the relevant abstract concepts. The choice, moreover, must be done in a principled manner. For if there were too much flexibility with respect to interpretive models—if the abstractness of concepts like force function or quantum Hamiltonian allow too much slack in the application of theories—then what reason would we have to hold that a theory receives corroboration or confirmation when it is successfully applied?

This is where bridge principles come in. Cartwright reinvents the positivist notion of a bridge principle (or correspondence rule). For her, bridge principles, which are part of the theory itself, use the stock of interpretive models to license concrete instantiations of the abstract concepts. These form the representations whose applications are licensed to convey warrant for the theory. Other applications are ad hoc and do not "give us really good reasons to believe in the truth of the theory" (195).

On this picture, theory application works something like this. We encounter some phenomenon. If all goes well, we can pick out some theory, which has, among its stock of interpretative models, one that makes possible the application of the abstract concepts of the theory to the phenomenon in question. The manner in which the model applies to the phenomenon is dictated by the bridge principles of the theory. If the modeling is empirically successful, we wind up with a representative model. (The theory, in turn, gains support.) Some phenomena, however, may not fit any single model. In these cases, we must construct a more complex representative model by coupling together the effects of several simple models. Somehow, this process is governed by capacities—somehow, although exactly how is not clear.

Earlier, we noted a tension between Cartwright's insistence on the plasticity of capacities and their purported function in regulating how we extend our theories beyond simple, ideal circumstances. The same tension occurs here. On the one hand, she holds that to represent what happens in concrete circumstances we need to go beyond theory and build a representative model and, on the other hand, she thinks that model construction is licensed by principles and interpretive models internal to the theory. The problem is that Cartwright is trying to capture two conflicting intuitions: the intuition that the task of identifying, for instance, a particular Hamiltonian for a concrete situation is a tricky and creative one, at best guided by theory, and the intuition that, in order for empirical success to confer warrant on a theory, model construction must be principled and not ad hoc.

In addition to this problem of coherence, one might wonder about Cartwright's assumption of a rigid dichotomy between "principled" and "ad hoc" applications of theory. It seems that different instances of modeling rely variously on their fidelity to theory, to data, and perhaps even fidelity to other considerations, such as analogy or previously successful modeling practice. Even applications of theory that might, in principle, be entirely theory driven are sometimes ruled out because of issues of mathematical tractability. For example, many fluid dynamical problems can be modeled in a principled way from the full-blown, compressible Navier-Stokes equations. Precious few of these models, however, result in a mathematically tractable set of equations. Finding working solutions in fluid dynamics almost always requires a motley mixture of theoretical principles, time-proven modeling techniques that are not necessarily theory driven, and experimental results and guesses.

So there seems to be a continuum between principled and ad hoc applications of theory. Moreover, the applications that lie far enough on the principled side to do the work that Cartwright requires seem a rather small lot—probably only toy models in textbooks fit the bill. (Sometimes, Cartwright seems to endorse this point herself. See her discussion of the Hamiltonian of the BCS model of superconductivity (199ff.).) She insists that the successful application of a theory to some phenomenon supports the theory only if every step in the construction of the representative model is fully licensed by components internal to the theory. This internalism looks like a residue of the vending-machine view and, in the end, she takes it mostly on faith that internalism will do justice to the variety of ways successful models support a theory.

We have ignored some interesting technical material in essays on probabilistic causality, economic models, and quantum physics. As in Cartwright's treatment of the more general issues of laws and models discussed here, on these topics, too, her ideas are subtle, original, and provocative. One may question whether the world is dappled and capacitated just in Cartwright's way. Beyond question, however, is that Cartwright always makes new philosophy and that hers is a way in which our discipline grows.

ERIC WINSBERG, MATHIAS FRISCH,
KAREN MERIKANGAS DARLING, and ARTHUR FINE
Northwestern University

NEW BOOKS

ADAMS, ROBERT MERRIHEW: *Finite and Infinite Goods: A Framework for Ethics.* New York: Oxford, 1999. xiv, 410 p. Cloth $45.00.

AGAMBEN, GIORGIO: *Potentialities: Collected Essays in Philosophy.* Stanford: University Press, 2000. x, 307 p. Cloth $55.00, paper $18.95.

ALSTON, WILLIAM P.: *Illocutionary Acts and Sentence Meaning.* Ithaca: Cornell, 2000. xiii, 327 p. Cloth $48.50.

BAKER, LYNNE RUDDER: *Persons and Bodies: A Constitution View.* New York: Cambridge, 2000. xii, 233 p. Cloth $49.95, paper $18.95.

BALABAN, ODED: *Plato and Protagoras: Truth and Relativism in Ancient Greek Philosophy.* Lanham, MD: Lexington, 1999. xx, 345 p. Cloth $75.00.

BANHAM, GARY: *Kant and the Ends of Aesthetics.* New York: St. Martin's, 2000. ix, 211 p. Cloth $59.95.

BELLANTONI, LISA: *Moral Progress: A Process Critique of MacIntyre.* Albany: SUNY, 2000. xii, 126 p. Paper $15.95.

BENARDETE, SETH: *Sacred Transgressions: A Reading of Sophocles'* Antigone. South Bend, IN: St. Augustine's, 1999. 162 p. Cloth $30.00.

BENSON, HUGH H.: *Socratic Wisdom: The Model of Knowledge in Plato's Early Dialogues.* New York: Oxford, 2000. xi, 292 p. Cloth $55.00.

BEN-ZE'EV, AARON: *The Subtlety of Emotions.* Cambridge: MIT, 2000. xiv, 588 p. Cloth $34.95.

BERLIN, ISAIAH: *The Power of Ideas.* HENRY HARDY, ed. Princeton: University Press, 2000. xv, 240 p. Cloth $24.95.

BEVERSLUIS, JOHN: *Cross-Examining Socrates: A Defense of the Interlocutors in Plato's Early Dialogues.* New York: Cambridge, 2000. xii, 416 p. Cloth $69.95.

BOGDAN, RADU J.: *Minding Minds: Evolving a Reflexive Mind by Interpreting Others.* Cambridge: MIT, 2000. xiii, 215 p. Cloth $35.00.

BOTTERILL, GEORGE, and PETER CARRUTHERS: *The Philosophy of Psychology.* New York: Cambridge, 1999. xii, 297 p. Cloth $57.95, paper $21.95.

BROOKE, JOHN, and GEOFFREY CANTOR: *Reconstructing Nature: The Engagement of Science and Religion.* Oxford: New York, 1999. xii, 367 p. Cloth $49.95, paper $24.95.

BUNGE, MARIO: *The Sociology-Philosophy Connection.* New Brunswick, NJ: Transaction, 1999. xix, 244 p. Cloth $29.95.

CARABINE, DEIRDRE: *John Scottus Eriugena.* New York: Oxford, 2000. xii, 131 p. Cloth $35.00, paper $16.95.

CARRETTE, JEREMY R.: *Foucault and Religion: Spiritual Corporality and Political Spirituality.* New York: Routledge, 2000. xvi, 215 p. Cloth $75.00, paper $24.95.

CARRIER, DAVID: *The Aesthetics of Comics.* University Park: Penn State UP, 2000. xii, 139 p. Cloth $29.95.

CARROLL, NOËL: *Philosophy of Art: A Contemporary Introduction.* New York: Routledge, 1999. xi, 273 p. Cloth $65.00, paper $18.99.

CASATI, ROBERTO, and ACHILLE C. VARZI: *Parts and Places: The Structures of Spatial Representation.* Cambridge: MIT, 1999. viii, 238 p. Cloth $35.00.

COLBURN, TIMOTHY R.: *Philosophy and Computer Science.* Armonk, NY: Sharpe, 2000. xi, 243 p. Cloth $58.95, paper $22.95.

COOK, JOHN W.: *Wittgenstein, Empiricism, and Language.* New York: Oxford, 2000. xvi, 224 p. Cloth $39.95.

CORBÍ, JOSEP E., and JOSEP L. PRADES: *Minds, Causes, and Mechanisms: A Case against Physicalism.* Malden, MA: Blackwell, 2000. xv, 205 p. Paper $29.95.

CORTENS, ANDREW JOSEPH: *Global Anti-Realism: A Metaphilosophical Inquiry.* Boulder: Westview, 2000. x, 174 p. Cloth $59.00.

COVER, J. A., and JOHN O'LEARY-HAWTHORNE: *Substance and Individuation in Leibniz.* New York: Cambridge, 1999. x, 307 p. Cloth $59.95.

CRARY, JONATHAN: *Suspensions of Perception: Attention, Spectacle, and Modern Culture.* Cambridge: MIT, 1999. xii, 397 p. Cloth $39.95.

CROSS, RICHARD. *Duns Scotus.* New York: Oxford, 1999. xxi, 250 p. Cloth $45.00, paper $19.95.

DIQUINZIO, PATRICE: *The Impossibility of Motherhood: Feminism, Individualism, and the Problem of Mothering.* New York: Routledge, 1999. xx, 275 p. Cloth $105.00, paper $29.00.

DRETSKE, FRED: *Perception, Knowledge, and Belief: Selected Essays.* New York: Oxford, 2000. xii, 284 p. Cloth $59.95, paper $19.95.

DUFRESNE, TODD: *Tales from the Freudian Crypt: The Death Drive in Text and Context.* Stanford: University Press, 2000. xiv, 229 p. Cloth $49.50, paper $19.95.

EDELMAN, GERALD M., and GIULIO TONONI: *A Universe of Consciousness: How Matter Becomes Imagination.* New York: Basic, 2000. xiii, 274 p. Cloth $27.50.

EKSTROM, LAURA WADDELL: *Free Will: A Philosophical Study.* Boulder: Westview, 2000. xi, 236 p. Cloth $65.00, paper $23.00.

ELKINS, JAMES: *Pictures of the Body: Pain and Metamorphosis.* Stanford: University Press, 1999. xvii, 347 p. Cloth $65.00, paper $24.95.

ELSTER, JON: *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints.* New York: Cambridge, 2000. xi, 308 p. Cloth $54.95, paper $19.95.

EVANS, G. R.: *Bernard of Clairvaux.* New York: Oxford, 2000. ix, 220 p. Paper $18.95.

FAIRFIELD, PAUL: *Moral Selfhood in the Liberal Tradition: The Politics of Individuality.* Toronto: University Press, 2000. viii, 278 p. Cloth $60.00.

FERREIRÓS, JOSÉ: *Labyrinth of Thought: A History of Set Theory and Its Role in Modern Mathematics.* Boston: Birkhäuser, 1999. xxi, 440 p. Cloth $130.00.

FEYERABEND, PAUL: *Conquest of Abundance: A Tale of Abstraction versus the Richness of Being.* BERT TERPSTRA, ed. Chicago: University Press, 1999. xviii, 285 p. Cloth $27.00.

FRAZER, ELIZABETH: *The Problems of Communitarian Politics: Unity and Conflict.* New York: Oxford, 2000. xi, 279 p. Cloth $70.00, paper $24.95.

FULLER, STEVE: *Thomas Kuhn: A Philosophical History of Our Times.* Chicago: University Press, 2000. xvii, 472 p. Cloth $35.00.

GAMWELL, FRANKLIN I.: *Democracy on Purpose: Justice and the Reality of God.* Washington D.C.: Georgetown, 2000. x, 366 p. Cloth $65.00.

GÄRDENFORS, PETER: *Conceptual Spaces: The Geometry of Thought.* Cambridge: MIT, 2000. x, 307 p. Cloth $40.00.

GASCHÉ, RODOLPHE: *Of Minimal Things: Studies on the Notion of Relation.* Stanford: University Press, 2000. x, 380 p. Cloth $55.00, paper $19.95.

GERENCSER, STEVEN ANTHONY: *The Skeptic's Oakeshott.* New York: St. Martin's, 2000. viii, 214 p. Cloth $49.95.

GIGERENZER, GERD, PETER M. TODD, and the ABC RESEARCH GROUP: *Simple Heuristics That Make Us Smart.* New York: Oxford, 1999. ix, 416 p. Cloth $35.00

GILL, JERRY H.: *The Tacit Mode: Michael Polanyi's Postmodern Philosophy.* Albany: SUNY, 2000. xiii, 192 p. Paper $17.95.

GOLDING, MARTIN P.: *Free Speech on Campus.* Lanham, MD: Rowman and Littlefield, 2000. ix, 118 p. Cloth $61.95, paper $17.95.

GOLDSTEIN, LAURENCE: *Clear and Queer Thinking: Wittgenstein's Development and His Relevance to Modern Thought.* Lanham, MD: Rowman and Littlefield, 1999. xii, 244 p. Cloth $63.00, paper $24.95.

GOODMAN, LENN E.: *Jewish and Islamic Philosophy: Crosspollinations in the Classic Age.* New Brunswick, NJ: Rutgers, 1999. xv, 256 p. Cloth $55.00.

GRAHAM, GORDON: *The Internet: A Philosophical Inquiry.* New York: Routledge, 1999. ix, 179 p. Paper $14.99.

GREEN, GARRETT: *Theology, Hermeneutics, and Imagination: The Crisis of Interpretation at the End of Modernity.* New York: Cambridge, 2000. xii, 229 p. Cloth $59.95.

GREENE, ROBERT: *The Death and Life of Philosophy.* South Bend, IN: St. Augustine's, 1999. x, 306 p. Cloth $30.00.

GUYER, PAUL: *Kant on Freedom, Law, and Happiness.* New York: Cambridge, 2000. xii, 440 p. Cloth $54.95, paper $19.95.

HACKER, P. M. S.: *Wittgenstein: Mind and Will, Volume 4 of an Analytical Commentary on the* Philosophical Investigations, *Part I: Essays.* Malden, MA: Blackwell, 2000. xxiii, 289 p. Paper $34.95; *Wittgenstein: Mind and Will, Volume 4 of an Analytical Commentary on the* Philosophical Investigations, *Part II: Exegesis §§428 - 693.* Malden, MA: Blackwell, 2000. xxii, 462 p. Paper $34.95.

HACKING, IAN.: *The Social Construction of What?* Cambridge: Harvard, 1999. x, 261 p. Cloth $29.95.

HALBERSTAM, MICHAEL: *Totalitarianism and the Modern Conception of Politics.* New Haven: Yale, 1999. x, 290 p. Cloth $27.50.

HALES, STEVEN D., and REX WELSHON: *Nietzsche's Perspectivism.* Urbana: Illinois UP, 2000. vii, 228 p. Cloth $40.00, paper $19.95.

HALL, RONALD L.: *The Human Embrace: The Love of Philosophy and the Philosophy of Love: Kierkegaard, Cavell, and Nussbaum.* University Park: Penn State UP, 2000. xii, 259 p. Cloth $35.00, paper $19.95.

HAMBURGER, JOSEPH. *John Stuart Mill on Liberty and Control.* Princeton: University Press, 1999. xx, 239 p. Cloth $35.00.

HAMPSHIRE, STUART: *Justice Is Conflict.* Princeton: University Press, 2000. xiii, 98 p. Cloth $18.95.

HARDCASTLE, VALERIE GRAY: *The Myth of Pain.* Cambridge: MIT, 1999. xv, 298 p. Cloth $35.00.

HARDWIG, JOHN: *"Is There a Duty to Die?" and Other Essays in Medical Ethics.* New York: Routledge, 2000. ix, 212 p. Cloth $85.00, paper $21.99.

HARRIS, ERROL E.: *Apocalypse and Paradigm: Science and Everyday Thinking.* Westport, CT: Praeger, 2000. xi, 154 p. Cloth $55.00.

HARRIS, GEORGE W.: *Agent-Centered Morality: An Aristotelian Alternative to Kantian Internalism.* Berkeley: California UP, 1999. xi, 434 p. Cloth $60.00.

HARRISON, S. J.: *Apuleius: A Latin Sophist.* New York: Oxford, 2000. viii, 281 p. Cloth $74.00.

HERDA, ELLEN A.: *Research Conversations and Narrative: A Critical Hermeneutic Orientation in Participatory Inquiry.* Westport, CT: Praeger 1999. xii,173 p. Cloth $59.95.

HIGGINS, KATHLEEN MARIE: *Comic Relief: Nietzsche's* Gay Science. New York: Oxford, 2000. xiii, 249 p. Cloth $45.00.

HINTIKKA, JAAKKO: *Inquiry as Inquiry: A Logic of Scientific Discovery.* Boston: Kluwer, 1999. xiii, 289 p. Cloth $132.00.

HIRSCHKOP, KEN: *Mikhail Bakhtin: An Aesthetic for Democracy.* New York: Oxford, 2000. xx, 332 p. Cloth $74.00, paper $24.95.

HODGES, MICHAEL, and JOHN LACHS: *Thinking in the Ruins: Wittgenstein and Santayana on Contingency.* Nashville: Vanderbilt, 2000. xiii, 128 p. Cloth $29.95.

HOLLEY, DAVID M.: *Self-Interest and Beyond.* St. Paul, MN: Paragon House, 1999. xiv, 236 p. Paper $16.95.

HUANG, SIU-CHI: *Essentials of Neo-Confucianism.* Westport, CT: Greenwood, 1999. xii, 261 p. Cloth $85.00.

HUME, DAVID: *A Treatise of Human Nature.* DAVID FATE NORTON and MARY J. NORTON, eds. New York: Oxford, 2000. ix, 622 p. Paper $14.95.

INGRAM, DAVID: *Group Rights: Reconciling Equality and Difference.* Lawrence: Kansas UP, 2000. ix, 323 p. Cloth $40.00, paper $17.95.

INWOOD, MICHAEL. *A Heidegger Dictionary.* Malden, MA: Blackwell, 1999. xvi, 283 p. Paper $24.95.

JAMMER, MAX: *Einstein and Religion: Physics and Theology.* Princeton: University Press, 1999. vii, 279 p. Cloth $22.95.

JESSEPH, DOUGLAS M.: *Squaring the Circle: The War between Hobbes and Wallis.* Chicago: University Press, 1999. xiv, 419 p. Cloth $80.00, paper $28.00.

JOHANSON, ARNOLD A.: *Principia Practica: The Logic of Practice.* Lanham, MD: UP of America, 2000. xx, 333 p. Cloth $57.50.

JOHNSON, DAVID: *Hume, Holism, and Miracles.* Ithaca: Cornell, 1999. ix, 106 p. Cloth $29.95.

JUARRERO, ALICIA: *Dynamics in Action: Intentional Behavior as a Complex System.* Cambridge: MIT, 1999. xi, 288 p. Cloth $40.00.

KADANE, JOSEPH B., MARK J. SCHERVISH, and TEDDY SEIDENFELD: *Rethinking the Foundations of Statistics.* New York: Cambridge, 1999. x, 388 p. Cloth $64.95, paper $24.95.

KAMITSUKA, DAVID G.: *Theology and Contemporary Culture: Liberation, Postliberal and Revisionary Perspectives.* New York: Cambridge, 1999. viii, 211 p. Cloth $54.95.

KATAYAMA, ERROL G.: *Aristotle on Artifacts: A Metaphysical Puzzle.* Albany: SUNY, 1999. xiii, 202 p. Paper $17.95.

KELLER, PIERRE: *Husserl and Heidegger on Human Experience.* New York: Cambridge, 1999. v, 261 p. Cloth $59.95.

KOHÁK, ERAZIM: *The Green Halo: A Bird's-Eye View of Ecological Ethics.* Chicago: Open Court, 2000. xiii, 209 p. Paper $24.95.

KRETZMANN, NORMAN: *The Metaphysics of Creation: Aquinas's Natural Theology in* Summa Contra Gentiles II. New York: Oxford, 1998. xii, 483 p. Cloth $65.00.

KUPPERMAN, JOEL J.: *Learning from Asian Philosophy.* New York: Oxford, 1999. viii, 208 p. Cloth $45.00, paper $19.95.

KUSCH, MARTIN: *Psychological Knowledge: A Social History and Philosophy.* New York: Routledge, 1999. xiii, 413 p. Cloth $99.00.

LACEY, HUGH: *Is Science Value Free? Values and Scientific Understanding.* New York: Routledge, 1999. xiv, 285 p. Cloth $85.00.

LANGAN, THOMAS: *Surviving the Age of Virtual Reality.* Columbia: Missouri UP, 2000. 1, 184 p. Cloth $29.95.

LEVINSON, MEIRA: *The Demands of Liberal Education.* New York: Oxford, 1999. ix, 237 p. Cloth $35.00.

LINGIS, ALPHONSO: *Dangerous Emotions.* Berkeley: California UP, 2000. 1, 195 p. Cloth $45.00, paper $16.95.

LLEWELYN, JOHN: *The Hypocritical Imagination: Between Kant and Levinas.* New York: Routledge, 2000. xiv, 274 p. Cloth $75.00, paper $25.99.

LOFTS, S. G.: *Ernst Cassirer: A "Repetition" of Modernity.* Albany: SUNY, 2000. xiv, 262 p. Cloth $73.50.

LOUDEN, ROBERT B.: *Kant's Impure Ethics: From Rational Beings to Human Beings.* New York: Oxford, 2000. xvii, 254 p. Cloth $45.00.

MALL, RAM ADHAR: *Intercultural Philosophy.* Lanham, MD: Rowman and Littlefield, 2000. xiii, 152 p. Cloth $62.00, paper $16.95.

MANDLE, JON: *What's Left of Liberalism? An Interpretation and Defense of Justice as Fairness.* Lanham, MD: Lexington Books, 2000. xi, 323 p. Cloth $77.00, paper $26.95.

MANN, WOLFGANG-RAINER: *The Discovery of Things: Aristotle's* Categories *and Their Context.* Princeton: University Press, 2000. xii, 231 p. Cloth $39.50.

MARTIN, MIKE W.: *Meaningful Work: Rethinking Professional Ethics.* New York: Oxford, 2000. xiii, 252 p. Cloth $35.00.

MASON, RICHARD: *Before Logic.* Albany: SUNY, 2000. 1, 153 p. Paper $16.95.

MAY, SIMON: *Nietzsche's Ethics and His War on 'Morality'.* New York: Oxford, 2000. xiv, 212 p. Cloth $45.00.

MCGILVRAY, JAMES: *Chomsky: Language, Mind, and Politics.* Malden, MA: Blackwell, 1999. xiii, 274 p. Cloth $54.95.

MILLER, DAVID: *Principles of Social Justice.* Cambridge: Harvard, 1999. xi, 337 p. Cloth $45.00.

MOI, TORIL: *What Is a Woman? And Other Essays.* New York: Oxford, 1999. xxv, 517 p. Cloth $35.00.

MORRISON, MARGARET: *Unifying Scientific Theories: Physical Concepts and Mathematical Structures.* New York: Cambridge, 2000. viii, 272 p. Cloth $59.95.

NEU, JEROME: *A Tear Is an Intellectual Thing: The Meanings of Emotion.* New York: Oxford, 2000. 1, 342 p. Cloth $45.00.

NOONAN, HAROLD W.: *Hume: On Knowledge.* New York: Routledge, 1999. ix, 221 p. Cloth $60.00, paper $14.99.

NOTTURNO, MARK AMADEUS: *Science and the Open Society: The Future of Karl Popper's Philosophy.* New York: Central European, 2000. xxvii, 287 p. Cloth $49.95, paper $23.95.

NOVAK, DAVID: *Conventional Rights: A Study in Jewish Political Theory.* Princeton: University Press, 2000. xiii, 240 p. Cloth $29.95.

NUSSBAUM, MARTHA C.: *Women and Human Development: The Capabilities Approach.* New York: Cambridge, 2000. xxi, 312 p. Cloth $24.95.

ODERBERG, DAVID S.: *Moral Theory: A Non-Consequentialist Approach.* Malden, MA: Blackwell, 2000. xiii, 197 p. Cloth $59.95, paper $27.95.

OGRODNICK, MARGARET: *Instinct and Intimacy: Political Philosophy and Autobiography in Rousseau.* Toronto: University Press, 1999. xi, 238 p. Cloth $50.00.

O'HAGAN, TIMOTHY: *Rousseau.* New York: Routledge, 1999. xii, 320 p. Cloth $85.00.

PALMER, JOHN A.: *Plato's Reception of Parmenides.* New York: Oxford, 1999. xiii, 294 p. Cloth $65.00.

PATTEN, ALAN: *Hegel's Idea of Freedom.* New York: Oxford, 1999. xiii, 216 p. Cloth $45.00.

PEARL, JUDEA: *Causality: Models, Reasoning, and Inference.* New York: Cambridge, 2000. xvi, 384 p. Cloth $39.95.

PENNOCK, ROBERT T.: *Tower of Babel: The Evidence against the New Creationism.* Cambridge: MIT, 2000. xix, 429 p. Cloth $18.95.

PIETERSMA, HENRY: *Phenomenological Epistemology.* New York: Oxford, 2000. xii, 204 p. Cloth $45.00.

PHILLIPS, D. Z.: *Recovering Religious Concepts: Closing Epistemic Divides.* New York: St. Martin's, 2000. xii, 272 p. Cloth $59.95.

PINKARD, TERRY: *Hegel: A Biography.* New York: Cambridge, 2000. xx, 780 p. Cloth $39.95.

PIPPIN, ROBERT B.: *Henry James and Modern Moral Life.* New York: Cambridge, 2000. xi, 193 p. Cloth $49.95.

PLANTINGA, ALVIN: *Warranted Christian Belief.* New York: Oxford, 2000. xx, 508 p. Cloth $60.00, paper $24.95.

POOLE, ROSS: *Nation and Identity.* New York: Routledge, 1999. xi, 209 p. Cloth $75.00, paper $24.99.

PSILLOS, STATHIS: *Scientific Realism: How Science Tracks Truth.* New York: Routledge, 1999. xxv, 341 p. Paper $27.99.

RAPOSA, MICHAEL L.: *Boredom and the Religious Imagination.* Charlottesville: Virginia UP, 1999. xiii, 199 p. Cloth $40.00, paper $13.95.

RAWLS, JOHN: *A Theory of Justice,* Revised Edition. Cambridge: Harvard, 1999. xxii, 538 p. Cloth $45.00, paper $22.00.

RÉE, JONATHAN: *I See a Voice: Deafness, Language, and the Senses: A Philosophical History.* New York: Henry Holt, 1999. xvi, 399 p. Cloth $27.50.

REEVE, C. D. C.: *Substantial Knowledge: Aristotle's* Metaphysics. Indianapolis: Hackett, 2000. xviii, 322 p. Cloth $34.95.

REITZ, CHARLES: *Art, Alienation, and the Humanities: A Critical Engagement with Herbert Marcuse.* Albany: SUNY, 2000. xvi, 336 p. Paper $17.95.

RISJORD, MARK W.: *Woodcutters and Witchcraft: Rationality and Interpretive Change in the Social Sciences.* Albany: SUNY, 2000. ix, 201 p. Paper $18.95.

ROEPER, P., and H. LEBLANC: *Probability Theory and Probability Semantics.* Toronto: University Press, 1999. ix, 240 p. Cloth $65.00.

ROSEMANN, PHILIPP W.: *Understanding Scholastic Thought with Foucault.* New York: St. Martin's, 1999. xiv, 263 p. Cloth $49.95.

ROSS, STEPHEN DAVID: *The Gift of Kinds: The Good in Abundance: An Ethic of the Earth.* Albany: SUNY, 1999. xvii, 323 p. Paper $24.95.

ROWLANDS, MARK: *The Body in Mind: Understanding Cognitive Processes.* New York: Cambridge, 1999. x, 270 p. Cloth $59.95.

RUDEBUSCH, GEORGE: *Socrates, Pleasure, and Value.* New York: Oxford, 1999. xvii, 169 p. Cloth $35.00.

SAVICKEY, BETH: *Wittgenstein's Art of Investigation.* New York: Routledge, 1999. ix, 266 p. Cloth $85.00.

SEIDEL, GEORGE J.: *Toward a Hermeneutics of Spirit.* Cranbury, NJ: Bucknell, 2000. 1, 154 p. Cloth $35.00.

SEREQUEBERHAN, TSENAY: *Our Heritage: The Past in the Present of African-American and African Existence.* Lanham, MD: Rowman and Littlefield, 2000. xiv, 103 p. Cloth $59.95, paper $17.95.

SHAPIRO, IAN: *Democratic Justice.* New Haven: Yale, 1999. xiii, 333 p. Cloth $29.95.

SHEETS-JOHNSTONE, MAXINE. *The Primacy of Movement.* Philadelphia: John Benjamins, 1999. xxxiii, 583 p. Paper $79.00.

SHEPHERDSON, CHARLES: *Vital Signs: Nature, Culture, Psychoanalysis.* New York: Routledge, 2000. xv, 256 p. Cloth $85.00, paper $23.99.

SHOOK, JOHN R.: *Dewey's Empirical Theory of Knowledge and Reality.* Nashville: Vanderbilt, 2000. ix, 316 p. Cloth $46.00, paper $22.95.

SHOPE, ROBERT K.: *The Nature of Meaningfulness: Representing, Powers, and Meaning.* Lanham, MD: Rowman and Littlefield, 1999. xii, 352 p. Cloth $65.00, paper $22.95.

SINGER, PETER: *A Darwinian Left: Politics, Evolution, and Cooperation.* New Haven: Yale, 1999. ix, 70 p. Cloth $9.95.

SMITH, DENNIS: *Zygmunt Bauman: Prophet of Postmodernity.* Malden, MA: Blackwell, 1999. xi, 250 p. Paper $24.95.

SMITH, TARA: *Viable Values: A Study of Life as the Root and Reward of Morality.* Rowman and Littlefield, 2000. ix, 205 p. Cloth $58.95, paper $17.95.

SOLOMON, ROBERT C.: *The Joy of Philosophy: Thinking Thin versus the Passionate Life.* New York: Oxford, 1999. xiii, 269 p. Cloth $29.95.

STALNAKER, ROBERT C.: *Context and Content: Essays on Intentionality in Speech and Thought.* New York: Oxford, 1999. x, 283 p. Cloth $65.00, paper $19.95.

STROUD, BARRY: *The Quest for Reality: Subjectivism and the Metaphysics of Colour.* New York: Oxford, 2000. xviii, 228 p. Cloth $29.95.

SUTTON, TIFFANY: *The Classification of Visual Art: A Philosophical Myth and Its History.* New York: Cambridge, 2000. xii, 184 p. Cloth $49.95.

TALLIS, RAYMOND: *On the Edge of Certainty: Philosophical Explorations.* New York: St. Martin's, 1999. xvii, 236 p. Cloth $45.00.

TORRETTI, ROBERTO: *The Philosophy of Physics.* New York: Cambridge, 1999. xvi, 512 p. Cloth $64.95, paper $23.95.

VAN CLEVE, JAMES. *Problems from Kant.* New York: Oxford, 2000. xii, 340 p. Cloth $45.00.

VILLA, DANA R.: *Politics, Philosophy, Terror: Essays on the Thought of Hannah Arendt.* Princeton: University Press, 1999. x, 266 p. Cloth $65.00, paper $17.95.

VRIES, HENT DE: *Philosophy and the Turn to Religion.* Baltimore: Johns Hopkins, 1999. xviii, 475 p. Cloth $55.00, paper $24.95.

WATKINS, JOHN: *Human Freedom after Darwin: A Critical Rationalist View.* Chicago: Open Court, 1999. xi, 348 p. Cloth $49.95, paper $24.95.

WEISS, PAUL: *Emphatics.* Nashville, TN: Vanderbilt, 2000. xii, 257 p. Cloth $34.95.

WEISSMAN, DAVID: *A Social Ontology.* New Haven: Yale, 2000. xv, 379 p. Cloth $40.00.

WERTZ, SPENCER K.: *Between Hume's Philosophy and History.* Lanham, MD: UP of America, 2000. xvi, 157 p. Cloth $49.00, paper $29.50.

WILSON, JOHN: *Learning to Love.* New York: St. Martin's, 2000. xiii, 301 p. Cloth $69.95, paper $19.95.

WOLLHEIM, RICHARD: *On the Emotions.* New Haven: Yale, 1999. xiii, 269 p. Cloth $25.00.

YOURGRAU, PALLE: *Gödel Meets Einstein: Time Travel in the Gödel Universe,* Expanded Edition. Chicago: Open Court, 1999. xxiv, 253 p. Paper $24.95.

ZUPANČIČ, ALENKA: *Ethics of the Real: Kant, Lacan.* New York: Verso, 2000. xiii, 266 p. Cloth $60.00, paper $22.00.

NEW BOOKS: ANTHOLOGIES

ALGRA, KEIMPE, JONATHAN BARNES, JAAP MANSFELD, and MALCOLM SCHOFIELD, eds.: *The Cambridge History of Hellenistic Philosophy.* New York: Cambridge, 1999. xix, 916 p. Cloth $135.00.

AMBROSIO, FRANCIS J., ed.: *The Question of Christian Philosophy Today.* New York: Fordham, 1999. xxvi, 366 p. Cloth $35.00, paper $17.50.

AMES, ROGER T., ed.: *The Aesthetic Turn: Reading Eliot Deutsch on Comparative Philosophy.* Chicago: Open Court, 2000. xiv, 225 p. Cloth $42.95.

BARNETT, STUART, ed.: *Hegel after Derrida.* New York: Routledge, 1998. x, 356 p. Cloth $85.00, paper $35.99.

BIANCHI, EMANUELA, ed.: *Is Feminist Philosophy Philosophy?* Evanston, IL: Northwestern, 1999. xxvii, 266 p. Cloth $69.95, paper $26.95.

BICCHIERI, CRISTINA, RICHARD JEFFREY, and BRIAN SKYRMS, eds.: *The Logic of Strategy.* New York: Oxford, 1999. xi, 195 p. Cloth $45.00.

BLACKBURN, SIMON, and KEITH SIMMONS, ed.: *Truth.* New York: Oxford, 1999. vi, 406 p. Cloth $65.00, paper $19.95.

BLOECHL, JEFFREY, ed: *The Face of the Other and the Trace of God.* New York: Fordham, 2000. xvi, 315 p. Cloth $39.95, paper $19.95.

BUBER, MARTIN: *On the Bible: Eighteen Studies by Martin Buber.* NAHUM N. GLATZER, ed. Syracuse: University Press, 2000. xxxii, 247 p. Paper $22.95.

BUCHANAN, IAN, ed.: *A Deleuzian Century?* Durham: Duke, 1999. iv, 274 p. Cloth $54.95, paper $18.95.

BURGESS-JACKSON, KEITH, ed.: *A Most Detestable Crime: New Philosophical Essays on Rape.* New York: Oxford, 1999. xv, 306 p. Cloth $45.00.

BUTTERFIELD, JEREMY, ed.: *The Arguments of Time.* New York: Oxford, 2000. xvi, 253 p. Cloth $49.95.

BUTTERFIELD, JEREMY, and CONSTANTINE PAGONIS, eds.: *From Physics to Philosophy.* New York: Cambridge, 2000. xv, 235 p. Cloth $59.95.

CAHN, STEVEN M., ed.: *Exploring Philosophy: An Introductory Anthology.* New York: Oxford, 2000. xii, 433 p. Cloth $35.00.

CAPUTO, JOHN D., and MICHAEL J. SCANLON, eds.: *God, the Gift, and Postmodernism.* Bloomington: Indiana UP, 1999. 1, 322 p. Cloth $49.95, paper $19.95.

CARD, CLAUDIA, ed.: *On Feminist Ethics and Politics.* Lawrence: Kansas UP, 1999. viii, 365 p. Cloth $45.00, paper $19.95.

CARROLL, NOËL: *Theories of Art Today.* Madison: Wisconsin UP, 2000. vi, 268 p. Cloth $55.00, paper $21.95.

CASTAÑEDA, HECTOR-NERI: *The Phenomeno-Logic of the I: Essays on Self-Consciousness.* JAMES G. HART and TOMIS KAPITAN, eds. Bloomington: Indiana UP, 1999. xi, 313 p. Cloth $45.00.

CLARKE, DESMOND M., and CHARLES JONES, eds.: *The Rights of Nations: Nations and Nationalism in a Changing World.* New York: St. Martin's, 1999. viii, 208 p. Cloth $59.95, paper $18.95.

COLLINGWOOD, R. G.: *The Principles of History and Other Writings in Philosophy of History.* W. H. DRAY and W. J. VAN DER DUSSEN, eds. New York: Oxford, 1999. lxxxvii, 293 p. Cloth $75.00.

COOPER, DAVID E., ed.: *Metaphysics: The Classic Readings.* Malden, MA: Blackwell, 2000. viii, 308 p. Paper $24.95.

CREATH, RICHARD, and JANE MAIENSCHEIN, eds.: *Biology and Epistemology.* New York: Cambridge, 1999. xviii, 295 p. Cloth $59.95, paper $19.95.

DAVION, VICTORIA, and CLARK WOLF, eds.: *The Idea of a Political Liberalism: Essays on Rawls.* Lanham, MD: Rowman and Littlefield, 2000. viii, 295 p. Cloth $62.95, paper $18.95.

DE CARO, MARIO, ed.: *Interpretations and Causes: New Perspectives on Donald Davidson's Philosophy.* Boston: Kluwer, 1999. vii, 259 p. Cloth $104.00.

DEWS, PETER, ed.: *Habermas: A Critical Reader.* Malden, MA: Blackwell, 1999. xvii, 348 p. Cloth $59.95, paper $24.95.

EGIDI, ROSARIA, ed.: *In Search of a New Humanism: The Philosophy of Georg Henrik von Wright.* Boston: Kluwer, 1999. vi, 238 p. Cloth $123.00.

FINE, GAIL, ed.: *Plato: Metaphysics and Epistemology, Volume 1.* New York: Oxford, 2000. viii, 514 p. Paper $19.95; *Plato: Ethics, Politics, Religion, and the Soul, Volume 2.* New York: Oxford, 2000. viii, 481 p. Paper $19.95.

GOLDBERG, KEN, ed.: *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet.* Cambridge: MIT, 2000. xix, 366 p. Cloth $35.00.

FISETTE, DENIS, ed.: *Consciousness and Intentionality: Models and Modalities of Attribution.* Boston: Kluwer, 1999. viii, 361 p. Cloth $140.00.

FRICKER, MIRANDA, and JENNIFER HORNSBY, eds.: *The Cambridge Companion to Feminism in Philosophy.* New York: Cambridge, 2000. xiii, 280 p. Cloth $54.95, paper $19.95.

GROSS, ALAN G., and ARTHUR E. WALZER, eds.: *Rereading Aristotle's* Rhetoric. Carbondale, IL: Southern Illinois UP, 2000. xi, 237 p. Cloth $34.95.

GENNARO, ROCCO J., and CHARLES HUENEMANN, eds.: *New Essays on the Rationalists.* New York: Oxford, 1999. xvii, 391 p. Cloth $60.00.

HAMEROFF, STUART R., ALFRED W. KASZNIAK, and DAVID J. CHALMERS, eds.: *Toward a Science of Consciousness III.* Cambridge: MIT, 2000. xx, 504 p. Paper $55.00.

HARDCASTLE, VALERIE GRAY, ed.: *Where Biology Meets Psychology: Philosophical Essays.* Cambridge: MIT, 1999. xii, 391 p. Cloth $55.00, paper $25.00.

HARVEY, STEVEN, ed. *The Medieval Hebrew Encyclopedias of Science and Philosophy.* Boston: Kluwer, 2000. xi, 547 p. Cloth $239.00.

HEMPEL, CARL G.: *Selected Philosophical Essays.* RICHARD JEFFREY, ed. New York: Cambridge, 2000. xii, 317 p. Cloth $54.95, paper $19.95.

HUMPHREYS, PAUL W., and JAMES H. FETZER, eds.: *The New Theory of Reference: Kripke, Marcus, and Its Origins.* Boston: Kluwer, 1998. xiii, 290 p. Cloth $105.00.

JACOBSON, ANNE JAAP, ed.: *Feminist Interpretations of David Hume.* University Park: Penn State UP, 2000. xi, 323 p. Cloth $55.00, paper $19.95.

JANAWAY, CHRISTOPHER, ed.: *The Cambridge Companion to Schopenhauer.* New York: Cambridge, 1999. xiv, 478 p. Cloth $59.95, paper $19.95.

JARVIE, IAN, and SANDRA PRALONG, eds.: *Popper's* Open Society *after Fifty Years: The Continuing Relevance of Karl Popper.* New York: Routledge, 1999. x, 217 p. Cloth $40.00.

JOHNSTON, DAVID, ed.: *Equality.* Indianapolis: Hackett, 2000. x, 267 p. Cloth $37.95, paper $16.95.

KEMAL, SALIM, and IVAN GASKELL, eds.: *Performance and Authenticity in the Arts.* New York: Oxford, 1999. x, 274 p. Cloth $59.95.

KEOWN, DAMIEN, ed.: *Buddhism and Abortion.* Honolulu: Hawai'i UP, 1999. xi, 222 p. Paper $22.95.

KLINE, T. C., III, and PHILIP J. IVANHOE, eds.: *Virtue, Nature, and Moral Agency in the* Xunzi. Indianapolis: Hackett, 2000. xxi, 268 p. Paper $16.95.

KÖGLER, HANS HERBERT, and KARSTEN R. STUEBER, eds.: *Empathy and Agency: The Problem of Understanding in the Human Sciences.* Boulder: Westview, 2000. ix, 318 p. Cloth $69.00, paper $26.00.

KUHSE, HELGA, and PETER SINGER, eds.: *Bioethics: An Anthology.* Malden, MA: Blackwell, 1999. ix, 600 p. Paper $39.95.

LAKATOS, IMRE, and PAUL FEYERABEND: *For and Against Method: Including Lakatos's Lectures on Scientific Method and the Lakatos-Feyerabend Correspondence.* MATTEO MOTTERLINI, ed. Chicago: University Press, 1999. xi, 451 p. Cloth $34.00.

LONG, A. A., ed.: *The Cambridge Companion to Early Greek Philosophy.* New York: Cambridge, 1999. xxxi, 427 p. Cloth $54.95, paper $19.95.

MACEDO, STEPHEN, ed.: *Deliberative Politics: Essays on Democracy and Disagreement.* New York: Oxford, 1999. xi, 289 p. Cloth $49.95, paper $19.95.

MACKENZIE, CATRIONA, and NATALIE STOLJAR, eds: *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self.* New York: Oxford, 2000. xi, 314 p. Cloth $49.95, paper $19.95.

MAGNANI, LORENZO, NANCY J. NERSESSIAN, and PAUL THAGARD, eds.: *Model-Based Reasoning in Scientific Discovery.* Boston: Kluwer, 1999. xi, 343 p. Cloth $135.00.

MANDER, W. J., ed.: *Anglo-American Idealism, 1865-1927.* Westport, CT.: Greenwood, 2000. vi, 232 p. Cloth $59.95.

MENDUS, SUSAN: *The Politics of Toleration in Modern Life.* Durham, NC: Duke, 2000. xii, 155 p. Cloth $49.95, paper $16.95.

NASR, SEYYED HOSSEIN, and MEHDI AMINRAZAVI, eds.: *An Anthology of Philosophy in Persia, Volume I.* New York: Oxford, 1999. xxxi, 434 p. Cloth $75.00.

NEWTON-SMITH, W. H., ed.: *A Companion to the Philosophy of Science.* Malden, MA: Blackwell, 2000. xvi, 576 p. Cloth $99.95.

NORTON, GLYN P., ed.: *The Cambridge History of Literary Criticism, Volume III: The Renaissance.* New York: Cambridge, 1999. xxiii, 758 p. Cloth $125.00.

O'HEAR, ANTHONY, ed.: *German Philosophy Since Kant.* New York: Cambridge, 1999. vi, 445 p. Paper $27.95.

OKIN, SUSAN MOLLER, et alia : *Is Multiculturalism Bad for Women?* Princeton: University Press, 1999. vi, 146 p. Cloth $29.95, paper $12.95.

PETITOT, JEAN, FRANCISCO J. VARELA, BERNARD PACHOUD, and JEAN-MICHEL ROY, eds.: *Naturalizing Phenomenology: Issues in Contemporary Phenomenology* *and Cognitive Science.* Stanford: University Press, 2000. xxi, 641 p. Cloth $75.00, paper $29.95.

PHILLIPS, D. Z., and TIMOTHY TESSIN, eds.: *Religion and Hume's Legacy.* New York: St. Martin's, 1999. xx, 282, p. Cloth $65.00.

PRESTON, JOHN, GONZALO MUNÉVAR, and DAVID LAMB, eds.: *The Worst Enemy of Science? Essays in Memory of Paul Feyerabend.* New York: Oxford, 2000. xvii, 171 p. Cloth $45.00.

PYLE, ANDREW, ed.: *Key Philosophers in Conversation: The* Cogito *Interviews.* New York: Routledge, 1999. ix, 256 p. Cloth $65.00, paper $20.99.

ROSEN, MICHAEL, JONATHAN WOLFF, and CATRIONA MCKINNON, eds.: *Political Thought.* New York: Oxford, 1999. xv, 442 p. Paper $21.95.

ROSENBERG, ALAN, JAMES R. WATSON, and DETLEF LINKE, eds.: *Contemporary Portrayals of Auschwitz: Philosophical Challenges.* New York: Prometheus, 2000. 1, 355 p. Cloth $69.95.

RUSSELL, BERTRAND: *Russell on Religion: Selections from the Writings of Bertrand Russell.* LOUIS GREENSPAN and STEFAN ANDERSON, eds. New York: Routledge, 1999. vii, 261 p. Cloth $65.00, paper $20.99.

SARGENT, ROSE-MARY, ed.: *Francis Bacon: Selected Philosophical Works.* Indianapolis: Hackett, 1999. xxxviii, 290 p. Cloth $34.95, paper $12.95.

SCOTT, CHARLES E., and JOHN SALLIS, eds.: *Interrogating the Tradition: Hermeneutics and the History of Philosophy.* Albany: SUNY, 2000. vii, 303 p. Paper $21.95.

SEDLEY, DAVID, ed.: *Oxford Studies in Ancient Philosophy, Volume XVII.* New York: Oxford, 1999. 1, 364 p. Cloth $60.00.

SHAPIRO, IAN, and CASIANO HACKER-CORDÓN, eds.: *Democracy's Edges.* New York: Cambridge, 1999. xiii, 297 p. Cloth $54.95, paper $19.95.

SCHRIFT, ALAN D., ed.: *Why Nietzsche Still? Reflections on Drama, Culture, and Politics.* Berkeley: California UP, 2000. xv, 309 p. Cloth $50.00, paper $19.95.

SOSA, ERNEST, JAEGWON KIM, and MATTHEW MCGRATH, eds.: *Epistemology: An Anthology.* Malden, MA: Blackwell, 2000. x, 590 p. Paper $44.95.

SPADE, PAUL VINCENT, ed.: *The Cambridge Companion to Ockham.* New York: Cambridge, 1999. xvii, 420 p. Cloth $54.95, paper $19.95.

STEEVES, H. PETER, ed.: *Animal Others: On Ethics, Ontology, and Animal Life.* Albany: SUNY, 1999. xiii, 294 p. Paper $17.95.

STERBA, JAMES P., ed.: *Ethics: Classical Western Texts in Feminist and Multicultural Perspectives.* New York: Oxford, 2000. v, 570 p. Paper $38.00.

ULLMANN-MARGALIT, EDNA, ed.: *Reasoning Practically.* New York: Oxford, 2000. ix, 188 p. Cloth $39.95.

WARREN, MARK E., ed.: *Democracy and Trust.* New York: Cambridge, 1999. xi, 370 p. Cloth $59.95, paper $22.95.

WESTPHAL, MEROLD, ed.: *Postmodern Philosophy and Christian Thought.* Bloomington: Indiana UP, 1999. vi, 291 p. Cloth $39.95, paper $19.95.

ŻEGLEŃ, URSULA M., ed.: *Donald Davidson: Truth, Meaning, and Knowledge.* New York: Routledge, 1999. ix, 186 p. Cloth $75.00.

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

crítica
REVISTA HISPANOAMERICANA DE FILOSOFÍA

Vol. XXXI / No. 93 / México, diciembre de 1999

Artículos

ALVIN I. GOLDMAN, Social Epistemology

ÁNGEL NEPOMUCENO FERNÁNDEZ, Tablas semánticas y metalógica (El caso de la lógica de segundo orden)

VÍCTOR SÁNCHEZ VALENCIA, Semántica aspectual y predicación secundaria

SILVIO PINTO, Wittgenstein on the Social Character of Language

Discusiones

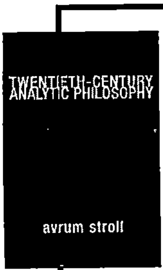
KAI NIELSEN, Moral Point of View Theories

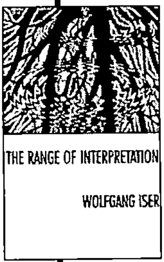
Notas bibliográficas

JOSÉ ALFREDO AMOR MONTAÑO, *Compacidad en la lógica de primer orden y su relación con el teorema de completud* [Atocha Aliseda]

CRÍTICA, Revista Hispanoamericana de Filosofía is published in April, August, and December. All correspondence should be addressed to *CRÍTICA*, Apartado 70-447, Coyoacán, 04510, D.F. México. e-mail: critica@filosoficas.unam.mx

---





Columbia's **philosophy.**

Twentieth-Century Analytic Philosophy
Avrum Stroll

Eminent philosopher and past Guggenheim fellow Avrum Stroll presents an authoritative and engaging history of analytic philosophy in this century, from Russell and Wittgenstein to Searle and Kripke.

304 pages • $32.50 cloth

The Range of Interpretation
Wolfgang Iser

"Iser's insightful and erudite book is a truly original contribution to the field, for a cultural-historical and anthropological exploration of the Western practices of interpretation has long been overdue."
—Mihai I. Spariosu, University of Georgia

280 pages • $25.00 cloth
THE WELLEK LIBRARY LECTURES

Sign up for bulletins of new books in your field at www.columbia.edu/cu/cup/catalog.

COLUMBIA UNIVERSITY PRESS
columbia.edu/cu/cup 800-944-8648

A THEORY OF SENTIENCE

AUSTEN CLARK

Clark offers a general account of the forms of mental representation that we call "sensory." Drawing on the findings of current neuroscience, he defends the hypothesis that the various modalities of sensation share a generic form that he calls "feature-placing."

2000 296 pp. $45.00

TWENTIETH CENTURY GERMAN PHILOSOPHY

PAUL GORNER

This book offers an historical and critical account of the important German philosophical movements and philosophers of the 20th century. Gorner introduces the reader to a principal representative of each movement, covering Husserl, Gadamar, Habermas, Apel, and Heideggar.

2000 168 pp. paper $14.95

Forthcoming!

EXPLAINING VALUE

and Other Essays in Moral Philosophy

GILBERT HARMAN

These thirteen essays represent a selection of the best of Harman's shorter writings in moral philosophy. They are divided into four sections, which focus in turn on moral relativism, values and valuing, character traits and virtue ethics, and ways of explaining aspects of morality.

August 2000 256 pp. paper $19.95 cloth $55.00

New in paperback!

THE FUTURE OF HUMAN REPRODUCTION

Ethics, Choice, and Regulation

Edited by JOHN HARRIS and SOREN HOLM

"The fourteen essays in this small volume provide an extremely useful and highly readable overview of the key moral, legal, and social themes raised by the new ways of making babies."—*British Medical Journal*

(Issues in Biomedical Ethics)

1999 (paper 2000) 264 pp. paper $19.95 cloth $45.00

REALISM REGAINED

An Exact Theory of Causation, Teleology, and the Mind

ROBERT KOONS

In this wide-ranging philosophical work, Koons takes on two powerful dogmas—anti-realism and materialism. In doing so, he develops an elegant metaphysical system that accounts for such phenomena as information, mental representation, mathematics and science, and the structure of spacetime.

2000 400 pp. $45.00

Forthcoming!

THINKING ABOUT MATHEMATICS

The Philosophy of Mathematics

STEWART SHAPIRO

This unique book looks at a range of philosophical issues and positions concerning mathematics, creating a sweeping introductory guide to the philosophy of mathematics.

August 2000 253 pp. paper $18.95

INALIENABLE RIGHTS

The Limits of Consent in Medicine and Law

TERRANCE MCCONNELL

This book explains what inalienable rights are and how they restrict the behavior of their possessors. McConnell develops compelling arguments to support the inalienability of the right to life, the right of conscience, and a competent person's right not to have medical treatment administered without consent.

2000 240 pp. $35.00

RESPECT, PLURALISM, AND JUSTICE

Kantian Perspectives

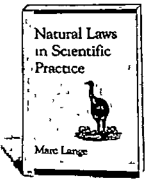
THOMAS E. HILL, Jr.

This series of essays sketches a broadly Kantian framework for moral deliberation, and then uses it to address important social and political issues. Hill shows how Kantian theory can be developed to deal with such issues as cultural diversity, political violence, and state coercion in a pluralistic society.

2000 280 pp. paper $19.95 cloth $60.00

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020.
www.oup.com



NATURAL LAWS IN SCIENTIFIC PRACTICE

MARC LANGE

Lange aims to develop a new account of the roles that laws of nature play in scientific reasoning (such as counterfactual conditionals, inductive projections, and scientific explanations) and what those roles imply about the very nature of natural laws.

2000 368 pp. $60.00

SELF AND OTHER

Philosophical Essays

J. N. MOHANTY

In the essays in this volume, Mohanty addresses contemporary questions of post-modernism without abandoning his fundamental stand on phenomenological method. He reveals a shift from an over-emphasis on identity in classical metaphysical thinking to an emphasis on differences without falling into the fogginess of post-modernism.

2000 204 pp. $19.95

MEANING, UNDERSTANDING, AND PRACTICE

Philosophical Essays

BARRY STROUD

In this collection, Stroud offers penetrating studies of meaning, understanding, necessity, and the intentionality of thought. His helpful introduction draws out the recurring themes he pursues and explains how his ideas and aims have developed over the years.

2000 232 pp. $35.00

UNDERSTANDING HUMAN KNOWLEDGE

Philosophical Essays

BARRY STROUD

Since the 1970s Stroud has been one of the most original contributors to the philosophical study of human knowledge. This volume presents the best of his essays in this area, in which he identifies the question that theories of knowledge are meant to answer, and the role scepticism plays in making sense of that question.

2000 256 pp. $35.00

THE PHILOSOPHY OF THE YOUNG KANT

MARTIN SCHÖNFELD

This intellectual biography of Immanuel Kant's early years makes an outstanding contribution to Kant scholarship. Schönfeld meticulously examines most of Kant's early works, summarizes their content, and exhibits their shortcomings and strengths.

2000 368 pp. $55.00

THE POSSIBILITY OF PRACTICAL REASON

J. DAVID VELLEMAN

This book explores the foundational questions of moral psychology. The eleven essays discuss such topics as the relation between value and practical reasoning, the foundations of decision theory, freedom of the will, shared intention, and the motivational role of the imagination.

2000 280 pp. paper $19.95 cloth $65.00

METAREPRESENTATIONS

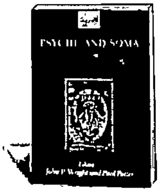
A Multidisciplinary Perspective

EDITED BY DAN SPERBER

This volume collects previously unpublished studies on metarepresentation by an interdisciplinary group of contributors, including Daniel Dennett, Alvin Goldman, Keith Lehrer, Leda Cosmides and John Tooby.

(Vancouver Studies in Cognitive Science 10)

2000 464 pp. paper $29.95 cloth $75.00



PSYCHE AND SOMA

Physicians and Metaphysicians on the Mind-Body Problem from Antiquity to Enlightenment

EDITED BY JOHN P. WRIGHT AND PAUL POTTER

This volume offers a multi-disciplinary exploration of our understanding of the human mind or soul and its relationship to the body through the course of more than two thousand years. Each of the especially commissioned chapters is written by a recognized expert.

2000 320 pp. $72.00

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020.
www.oup.com

Young Epistemologist Prize

The **Rutgers Epistemology Conference** announces the establishment of an annual prize for the best essay submitted to the Conference by a person with a Ph.D. obtained by the time of submission but not earlier than ten (10) years prior to the date of the conference. (Thus, for the Conference on April 27/28th of 2001, the Ph.D. must not have been awarded prior to April, 1991.) The prize winning essay will be presented at the annual conference and will be published (along with the invited papers) in a special issue of *Philosophy and Phenomenological Research.* The winner of the prize will receive an award of $1,000 plus all travel and lodging expenses connected with the Conference.

Details

The essay may be in any area of epistemology. The essay *must* be limited to 5,000 words (not including footnotes and bibliography) and submitted on a PC-formatted disk along with three (3) printed copies. The essays will be judged blindly and, thus, the three printed copies (but not the disk copy) must be prepared in such a way that the author's identity is completely masked. *Only* papers submitted in that fashion will be considered. Papers *must* be postmarked by December 1, 2000. The winner of the prize will be announced by February 1, 2001. By submitting the essay, the author agrees to have it posted on the Rutgers Epistemology Conference Webpage and published in *Philosophy and Phenomenological Research.*

Please send all papers (along with disk) to:

The Rutgers Epistemology Conference
Philosophy Department
Rutgers University
26 Nichol Ave.
New Brunswick, NJ 08901-2882

Invited speakers for the **2001 Rutgers Epistemology Conference** will be: Stewart Cohen, Peter Railton and Crispin Wright. For more information about the Conference and the Young Epistemologist Prize, please refer to:

http://philosophy.rutgers.edu/epistemology2001/

Best in Philosophy from Cambridge

The Concept of the Gene in Development and Evolution

Historical and Epistemological Perspectives

***Peter Beurton, Raphael Falk,** and **Hans-Jörg Rheinberger,** Editors*

This book is unique in that it is the first interdisciplinary volume, written by philosophers, historians, and working scientists, solely devoted to the quest for the gene.

Cambridge Studies in Philosophy and Biology

0-521-77187-0	Hardback	$59.95

From Chance to Choice

Genetics and Justice

Allen Buchanan, Dan Brock, Norman Daniels,** and **Daniel Wikler

This important book, written by four internationally renowned bioethicists, is the first systematic treatment of the fundamental ethical issues underlying the application of genetic technologies to human beings.

0-521-66001-7	Hardback	$29.95

Putting Skeptics in Their Place

The Nature of Skeptical Arguments and Their Role in Philosophical Inquiry

John Greco

This book is about the nature of skeptical arguments and their role in philosophical inquiry.

Cambridge Studies in Philosophy

0-521-77263-X	Hardback	$54.95

Edward Said and the Religious Effects of Culture

William Hart

This book provides a distinctive account of Edward Said's critique of modern culture by highlighting the religion-secularism distinction on which it is predicated.

Cambridge Studies in Religion and Critical Thought 8

0-521-77052-1	Hardback	$54.95
0-521-77810-7	Paperback	$19.95

The State and International Relations

John Hobson

The book offers readers a unique introduction to state theory in International Relations, and will be of interest to students and scholars of sociology and politics, as well as International Relations.

Themes in International Relations

0-521-64354-6	Hardback	$52.95
0-521-64391-0	Paperback	$18.95

Immigration as a Democratic Challenge

Citizenship and Inclusion in Germany and the United States

Ruth Rubio-Marín

At present, immigrants to most Western democracies must satisfy a range of conditions before achieving citizenship. This book argues that this is unjust and undemocratic, and that there should be a time threshold after which immigrants should either be granted full citizenship rights, or should be awarded nationality automatically, without any conditions.

0-521-77152-8	Hardback	$57.95
0-521-77770-4	Paperback	$21.95

The Classification of Visual Art

A Philosophical Myth and Its History

Tiffany Sutton

This book is an important and original contribution to the philosophy of art that bridges the disciplines of philosophy and art.

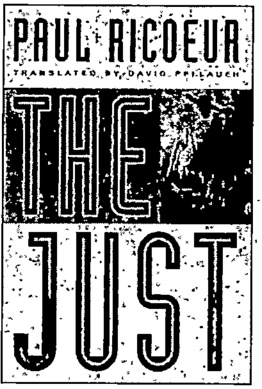
0-521-77236-2	Hardback	$49.95

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

40 West 20th St., N.Y., NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cup.org
AmEx/MasterCard/VISA accepted.
Prices subject to change.

New from Chicago



The Just

Paul Ricoeur

Translated by David Pellauer

Noted French philosopher Paul Ricoeur revisits the work of Plato, Aristotle, and Kant in his engagements with contemporary thinkers. His thought ranges from conceptual analysis, to the theory of law, and finally to the act of judging, as he explores the ideas of sanction, rehabilitation, pardon, and the status of conscience in relation to the demands of the law.

Cloth $20.00

The Argument of the Action

Essays on Greek Poetry and Philosophy

Seth Benardete

This volume brings together Seth Benardete's studies of Hesiod's *Theogony,* Homer's *Iliad,* and Greek tragedy, of eleven Platonic dialogues, and Aristotle's *Metaphysics.* These essays, some never before published, others difficult to find, span four decades of his work and document its impressive range. Benardete's original interpretations are the fruits of this discovery of the "argument of the action."

Cloth $39.00

On Tyranny

Leo Strauss

Edited by Victor Gourevitch and Michael S. Roth

Revised and Expanded Edition, Including the Strauss-Kojève Correspondence

"*On Tyranny* is a complex and stimulating book with its 'parallel dialogue' made all the more striking since both participants take such unusual, highly provocative positions, and so force readers to face substantial problems in what are often wholly unfamiliar, even shocking ways."
—Robert Pippin, *History and Theory*

Paper $17.00

Philosophy of Illumination

Shihāb al-Dīn Suhrawardī

Translated by John Walbridge and Hossein Ziai

Shihāb al-Dīn al-Suhrawardī was born around 1154, probably in northwestern Iran. Spurred by a dream in which Aristotle appeared to him, he rejected the Avicennan Peripatetic philosophy of his youth and undertook the task of reviving the philosophical tradition of the "Ancients."

Cloth $29.95

The University of Chicago Press

5801 South Ellis Avenue, Chicago IL 60637
www.press.uchicago.edu

29\8\

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NUMBER 8
AUGUST 2000

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

page
421 *Utilitarianism, Integrity, and Partiality* Elizabeth Ashford
440 *Surprise, Self-Knowledge, and Commonality* Frederic Schick

COMMENTS AND CRITICISM
454 *Fodor and Lepore on Meaning Similarity and Compositionality* Barbara Abbott
456 *Leveling the Playing Field between Mind and Machine: A Reply to McCall* Alexander George and Daniel J. Velleman
462 *What Gödel's Incompleteness Result Does and Does Not Show* Haim Gaifman

BOOK REVIEWS
471 Martha Nussbaum: *Sex and Social Justice* Miranda Fricker

476 NEW BOOKS: TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc. ISSN 0022–362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NO. 8, AUGUST 2000

UTILITARIANISM, INTEGRITY, AND PARTIALITY*

Bernard Williams's[1] integrity objection against utilitarianism has made a very influential contribution to the view that utilitarianism is so demanding that it cannot be a serious option. Utilitarians, on the other hand, have generally denied that a suitably sophisticated version of utilitarianism is incompatible with agents' integrity. I argue here that, if we examine what a valuable conception of integrity consists in, we can see that it actually commits us, in the current state of the world, to extremely demanding moral obligations, on *any* plausible account of our moral obligations, including Williams's own. I then argue, however, that any such account of these obligations has difficulty in providing a rationale for how a fundamental conflict between them and agents' pursuit of their personal projects can be avoided. I conclude that it is, in fact, a strength of utilitarianism that it acknowledges that this conflict cannot be resolved and makes explicit the extent to which our integrity is currently compromised. I lastly argue that there is a practically realizable state of the world in which utilitarian moral obligations would not seriously conflict with agents' pursuit of their personal projects.

I. A SUMMARY OF WILLIAMS'S INTEGRITY OBJECTION

Williams's integrity objection is directed at act utilitarianism, according to which the right action is the action that maximizes overall well-being. He claims that utilitarianism requires the agent to aban-

---

* I am grateful to Simon Blackburn, Roger Crisp, William Lycan, Derek Parfit, Geoffrey Sayre-McCord, Galen Strawson, and Christopher Taylor for their extremely helpful comments on previous drafts.

[1] I focus on the objection as he presents it in *Utilitarianism For and Against* (New York: Cambridge, 1973).

0022-362X/00/9708/421–39

don his projects whenever doing so will enable him to promote the most overall well-being, and to regard his projects as merely "one set of satisfactions among those which he may be able to assist from where he happens to be" (*ibid.*, p. 115). Williams then argues that, while it may be acceptable that morality requires the agent to abandon certain projects, it is unreasonable for morality to demand that the agent abandons those projects "with which [he] is more deeply and extensively involved and identified," and which he may "take seriously at the deepest level, as being what his life is about" (*ibid.*, p. 116). These are projects adherence to which the agent sees as constitutive of who he is—Williams terms them *ground projects.* They presumably bring together the agent's narrative identity and provide him with a sense of coherence across time.

Williams argues that the demand that an agent be prepared to set aside such ground projects in order to promote greater overall utility is "in the most literal sense, an attack on his integrity" (*ibid.*, p. 117). By "integrity" in the literal sense, I take it that Williams is referring to its classical meaning of "wholeness," and is using it to denote the agent's unified sense of self. Giving up ground projects with which the agent is identified would result in a degree of psychological fragmentation.

Among identity-conferring commitments are moral commitments. Williams points out that utilitarianism may ask the agent to act against his moral feelings even when they are central to his moral self-conception. Williams then argues that to regard such feelings "as happenings outside one's moral self, is to lose a sense of one's moral identity; to lose, in the most literal way, one's integrity" (*ibid.*, p. 104). Again, the argument rests on the importance of integrity conceived as the agent's sense of self, though it is here concerned specifically with the agent's *moral* self-conception.

Williams gives two examples to illustrate the way in which utilitarian requirements may alienate agents from their moral self-conceptions. George, a chemist, is required to take a job in a chemical warfare factory, despite his opposition to chemical warfare, to prevent a more talented and less scrupulous chemist from taking the job instead and pursuing it with more zeal. Jim is required to shoot one Indian in order that the lives of nineteen others will be spared.

Williams seems to assume that George and Jim are not utilitarians. He argues that their performing the utilitarian right action will involve their acting against their conscience, as the result of which their moral identity will be threatened. A utilitarian response to the examples and George and Jim, however, is that they should change the moral beliefs that underlie their current moral self-conceptions;

they should come to see the right action as the one that minimizes suffering or death. Performing the utilitarian right action would then no longer involve their acting against their conscience.

One interpretation of Williams's integrity objection is that it involves the claim that those moral feelings which are central to the agent's current moral self-conception should be taken as given, and as presenting constraints on what a moral theory can legitimately demand. This precludes the possibility of a moral theory's claiming that these moral feelings are mistaken.

Williams's objection, on this first interpretation, comprises two claims: first, the claim that it is a condition on the acceptability of a moral theory that it does not require agents to act in a way that contravenes their present self-conception, whatever that self-conception should consist in; and, second, the claim that utilitarianism undermines those commitments, moral and personal, adherence to which the agent sees as constitutive of who he is. I shall discuss this interpretation in the following section.

II. FIRST INTERPRETATION: THE THREAT TO THE AGENT'S PRESENT SELF-CONCEPTION

On this first interpretation, then, Williams's objection is that the cost to agents of the undermining of their integrity, conceived purely as their current unified self-conception, is unacceptably high, and that utilitarian requirements may impose such a cost. I shall argue that integrity on this conception cannot plausibly outweigh moral obligations, and, moreover, that is not the kind of integrity that agents take to be valuable. My argument will focus on the claim that agents' moral self-conceptions provide base-line constraints on the acceptability of moral requirements. I shall argue that, conversely, moral requirements provide constraints on the acceptability of agents' moral self-conceptions.

The example of slavery demonstrates that agents' moral self-conceptions can conflict with extremely strong moral obligations and, furthermore, that this conflict need not be confined to rare individuals, but may affect a whole society. When a form of immorality involves living by accepted social norms, the moral defects and inconsistencies of the way of life may not be apparent to members of the society if they fail to question those norms. They are able to give a coherent narrative to their life that enables them to retain a self-conception as morally upright. They are aided in this by the fact that agents' sense of identity is largely formed by the way they are viewed by those with whom they interact. Slave owners, for example, were able to retain their self-conception as morally decent despite owning slaves, because they were inculcated with the view that slavery was morally justifiable and, in addition, because they could gain positive recognition in the society that condoned the practice. If, on the other hand, an agent began to question the morality of owning slaves, then even if he attempted to oppose it he might feel morally compromised just by living in a society based on a slave-run economy.

Clearly, any plausible account of moral obligations must hold that slave owners should have questioned and opposed their way of life and the norms to which they subscribed, even at the cost of alienation from what were previously deeply held commitments and moral self-conceptions. The example thus demonstrates that agents' integrity in the sense of their current unified self-conception can be incompatible with undoubtedly overriding moral demands. It therefore cannot be a plausible demand on a moral theory that it be compatible with such integrity.

I conclude that in order to be forceful, the integrity objection cannot be grounded on the value of integrity conceived merely as the agent's current unified self-conception. It must appeal to what I will call *objective integrity*. For the agent to have objective integrity, her self-conception must be grounded in reality: it must not be based on her being seriously deceived either about empirical facts or about the moral obligations she actually has. In particular, her self-conception as being morally decent must be grounded in her leading a genuinely morally decent life. By contrast, the mere possession of a coherent self-conception, however mistaken, can be called *subjective integrity*.

A further reason why a forceful version of the integrity objection should appeal to objective integrity is that it is a considerably more plausible candidate than subjective integrity for what agents actually take to be valuable. The integrity objection appeals to the value of integrity, and hence the immense cost to agents of losing it. Its cogency will therefore depend on its appealing to the kind of integrity that agents in fact value.

Williams emphasizes the role played by agents' adherence to their central projects in grounding their sense of meaning in their lives. Robert Nozick's[2] experience-machine thought experiment indicates that the kind of sense of meaning we value is one grounded in reality, through our actually accomplishing something with our lives; it is not merely our having the subjective experience of leading a meaningful life which we value. In order for the agent to achieve genuine accomplishment, she must not be seriously deceived about the empirical facts relevant to her actions. But in addition, since a central component of genuine accomplishment is leading a life that is actually morally worthwhile, she must not be seriously deceived about her moral obligations. This indicates that we value having objective integrity, as opposed to merely having a self-conception according to which we are leading a worthwhile life.

---

[2] *Anarchy, State, and Utopia* (New York: Basic, 1974), pp. 42–45.

Another related reason why we value objective integrity is that the sense of self-worth which it confers is grounded in attributes the agent actually has. By contrast, subjective integrity is grounded only in the agent's retaining the belief that she possesses such attributes. If her self-conception does not correspond to reality, then she will need to find external means of retaining it. One way of doing so is by gaining positive recognition from others. This will necessitate the agent's dependency on others for her sense of self.

Objective integrity may lead to inner turbulence, since it involves an uncompromising concern for and pursuit of the truth about matters of importance, especially morality, through autonomous moral appraisal of oneself and of one's society's standards. Nevertheless, the sense of identity which it grounds is genuine, stable, and self-reliant.

I conclude that to be plausible the objection should be interpreted as claiming that utilitarianism undermines agents' objective integrity. I shall consider two versions of this claim: first, that utilitarianism undermines objective integrity in the *current* state of the world; second, that utilitarianism undermines objective integrity in *any* practically realizable state of the world.

III. SECOND INTERPRETATION: THE THREAT TO OBJECTIVE INTEGRITY IN THE CURRENT STATE OF THE WORLD

In order to assess the force of the objection that agents' current utilitarian moral obligations undermine their objective integrity, I shall first discuss the circumstances in which objective integrity is possible. I shall then argue that in the current state of the world, it is plausible and appropriate to claim that agents' objective integrity is inevitably threatened.

A. The conditions for objective integrity. I have argued that in order for us to have objective integrity, our moral self-conception must be grounded in our actually leading a morally decent life. This requires that we abide by our moral commitments and that these commitments stem from the moral obligations we actually have.

Both subjective and objective integrity also require the agent's adhering to her personal commitments, since, as Williams argues, these are centrally constitutive of any agent's identity and sense of the meaningfulness of her life. Personal commitments include prudential projects and commitments to particular individuals. Prudential projects are central to agents' personal fulfillment and, as Williams says, give them an interest in living their life. Commitments to particular persons, with whom the agent has a deep personal relationship, are genuinely other-regarding, since they involve a direct concern for others' interests. But they are, unlike impartial moral commitments, focused on those with whom the agent identifies and so are self-referential.

Impartial moral obligations are what generally differentiate subjective and objective integrity. Failure to comply with self-referential moral commitments will affect every agent's moral self-image, since agents' identities are automatically bound up in these commitments. By contrast, failure to comply with impartial moral obligations may have no impact on the agent's moral self-image and so may be compatible with subjective integrity. (The main reason for this is the impact of psychological distancing on agents' moral reactions, as the example of slave owners illustrates. Nineteenth-century slave owners, for example, were distanced from the suffering of slaves by viewing them as subhuman. We can assume that their sense of moral self-worth was entirely dependent on their interactions with their family, friends, and colleagues.) Such failure is, however, incompatible with leading a genuinely morally decent life and so having objective integrity.

Since objective integrity is thus grounded in the agent's adherence to both impartial moral commitments and to her personal projects and commitments, it can be achieved only by harmoniously realizing these two kinds of commitments. The concept of the good life (which roughly corresponds to Plato and Aristotle's concept of *eudaimonia*) sheds light on how the two kinds of commitments can be integrated. The good life has two central aspects: first, a good life is one that is genuinely meaningful as it is part of what we value in our lives that they have point and substance; second, it is personally fulfilling. Each of these aspects of the good life involves the pursuit of both impartial and personal projects. The agent's engagement in her personal projects is, as Williams says, central to her sense of meaningfulness. But in addition, a genuinely meaningful life is one that is morally worthwhile. Personal fulfillment is itself a prudential value. Nevertheless, it is largely realized by activities that bring a feeling of accomplishment and, as James Griffin[3] argues, the value of accomplishment centrally includes the agent's contributing to the well-being of others.

---

[3] *Value Judgement* (New York: Oxford, 1996), p. 70.

The good life can be seen as a kind of synoptic scale, against which moral and prudential projects can be weighed up and harmoniously integrated. For example, taking up an expensive hobby might be very personally fulfilling, but if the agent judges it to be incompatible with morally sufficient donation to aid agencies, then she will conclude that the hobby would not be part of an overall good life. Thus, a weighty moral commitment outweighs a relatively trivial prudential project. The difficult conflict arises, however, when the two components both essential to a good life fundamentally and irresolvably clash and so cannot be integrated into an overall good life. It is in such circumstances that agents' objective integrity is threatened.

B. The threat posed by utilitarianism to agents' objective integrity. In the current state of the world, agents' important utilitarian moral obligations may fundamentally clash with their pursuit of their personal projects. Furthermore, unlike most utilitarians who reject the over-demandingness charge against utilitarianism,[4] I agree with Williams that this clash threatens agents' integrity.

The plausibility of this can be illustrated by examining Roger Crisp's[5] example of the three lives open to Anna. In the first life, she has a lucrative career and flourishing personal relationships, but does nothing to alleviate the suffering of distant strangers. In the second life, she makes sizeable donations of time and money to aid agencies, but also allows herself enough leisure time to develop and maintain flourishing personal relationships, and enough time and money to pursue a range of projects. In the third life, she devotes herself solely to aid work, allowing herself no leisure time for engaging in deep personal relationships, and spending all her income beyond what is needed to keep herself nourished and healthy on helping others' urgent medical needs.

According to utilitarianism, although the third life may deprive Anna of central components of a flourishing life, it is nevertheless the right life to choose, because the loss to her is morally outweighed by the extreme suffering of several persons which the third life enables her to alleviate. Intuitively, however, it seems clear that she also has an extremely strong prudential reason to choose the second life over the third. A life such as the third, which involves giving up close relationships and other personal projects, could feel impoverished in many ways. Moreover, the depth of the personal sacrifice that utilitarianism would require of her can be brought out particularly saliently by considering that the third life could undermine her integrity.

---

[4] Shelly Kagan, for example, argues that the agent who fully complies with the impartial moral obligations to perform the acts which have the best overall consequences is the only agent who has genuine integrity—*The Limits of Morality* (New York: Oxford, 1989), pp. 390-92.

[5] "The Dualism of Practical Reason," *Proceedings of the Aristotelian Society*, XCVI (1996): 53-73, here p. 62.

There are some personal commitments an agent can give up and replace with new ones, without psychological dislocation. There are other commitments, however, including certain deep personal relationships, which may be life-long, and may form a permanent part of our identities and of the meaningfulness of our lives. If Anna had the personality and calling of a saint such as Mother Theresa, she might find the third life more personally fulfilling and enriching than the second one. If the third life involved her giving up a marriage and other close personal relationships, however, it is likely that she would find it self-fracturing.

I therefore agree with Williams that there may be an irresolvable conflict between utilitarian moral commitments and commitments of partiality, and that this may render objective integrity impossible. As I shall now argue, however, this problem is not unique to utilitarianism. To begin with, according to Williams's own account of our current moral obligations, there is an irresolvable conflict between such obligations and agents' pursuit of their personal projects.

C. An internal argument against Williams. According to Williams's[6] account of moral obligation, we have stringent moral obligations to drop our own concerns and help others in emergency situations. He argues that the notion of obligation (which prescribes that we *must* do or avoid doing a particular action whether we happen to want to or not) applies only in two contexts: first, negatively, in the injunction against killing and the like, where this injunction serves particularly important interests; and, second, positively, in the injunction to respond to an emergency, where our general disposition of concern for people's vital interests is focused by the immediate plight of particular individuals.

In response to the moralists who "say that if we regard immediacy or physical nearness as relevant we must be failing in rationality or imagination," he argues as follows:

> ...such moralists are wrong, at least in trying to base their challenge on the structure of obligation. Of course this point does not dispose of the challenge itself. We should be more concerned about the sufferings of people elsewhere. But a correct understanding of what obligation is will make it clearer how we should start thinking about the challenge. We

---

[6] *Ethics and the Limits of Philosophy* (Cambridge: Harvard, 1985), pp. 185-86.

> should not banish the category of immediacy, but we must consider what for us, in the modern world, should properly count as immediacy (*ibid.*, p. 186).

I interpret this as meaning that the objection does not threaten his account of positive obligation, as requiring immediacy to produce an obligation from the agent's underlying general concern, but rather that our conception of immediacy is what should come under scrutiny (presumably in the light of our modern media coverage and ability to help the vital interests of persons in distant countries). This implies that, on his account, the suffering of those in distant countries can come under the category of the immediate. If so, then I take it that by the term 'immediacy' he is referring not to geographical proximity, but to the imminent danger to particular individuals whom we are in a position to help.

If, on the other hand, he does maintain that immediacy involves physical proximity, then unless he can show that the geographical location of the threat to vital interests is morally relevant, agents' compliance with such obligations would ground only their subjective integrity and not their objective integrity. Physical proximity clearly has a considerable impact on our moral responses and on the impact of our failure to help on our moral self-conceptions, but this may be because of morally irrelevant psychological distancing. Moreover, Williams's own account of the ultimate foundation of the positive obligation to respond to an emergency (as well as the negative obligation not to kill others) appeals to a consideration that is unaffected by geographical location, namely, the urgency of the interests at stake:

> The obligations...are both based ultimately on one conception, that each person has a life to lead. People need help but (unless they are very young, very old, or severely handicapped) not all the time. All the time they need not to be killed, assaulted, or arbitrarily interfered with (*ibid.*, p. 186).

Williams's account of the obligation to respond to emergency situations therefore seems incompatible with taking immediacy to involve proximity.

If immediacy does not involve physical proximity, however, it is clear that Williams has drastically underestimated the extent of agents' positive obligations to give help in emergencies. In the lines I have just quoted, Williams assumes that while negative obligations are persistently operative, positive obligations by contrast arise only occasionally. This would mean that agents' ground projects were minimally threatened by the obligation to help persons' vital interests. His assumption is based on the claim that, with only some exceptions, such as the very old and young and the severely disabled, persons' vital interests are very rarely at risk. But once we include all those round the world whose chronic poverty severely impairs their ability to realize the most basic of their interests, his assumption that the threat to persons' vital interests arises only occasionally can be seen to be mistaken. And if the criterion for the positive obligation to give help in an emergency is based on the threat to persons' vital interests, then there is no rationale for limiting its scope to the old, young, and severely disabled. Given the vast number of those whose vital interests are threatened by malnutrition and disease, the obligation to help them will be frequently incompatible with agents' devoting money and energy to their own ground projects.

This suggests that the source of the extreme demandingness of morality is that the current state of the world is a constant emergency situation; there are continually persons whose vital interests are threatened and, given modern communications, the relatively well-off are continually able to help them. Because of the scale of the emergency, the moral requirement to help others' vital interests is so demanding as to threaten agents' personal projects. This formulation of the problem is one Williams should accept. The term 'emergency' emphasizes the immediacy of the threat to persons' vital interests, which should focus our general disposition of concern for others' vital interests (which is why the formulation has rhetorical force). Applying the term 'emergency' to the current state of the world follows his suggestion that we seriously reconsider what in the modern world we should properly count as immediacy.

The reason Williams's account of obligation threatens agents' integrity is that, given his account of the rationale for the obligation to give help in an emergency situation, the stopping point he tries to draw for the expansion of concern is arbitrary; and any plausible account of morality must face this problem of avoiding setting arbitrary boundaries on the sphere of moral concern. For example, common-sense morality maintains that in emergency situations, agents have a positive obligation to prevent people from dying. It maintains that the moral obligation to avoid actively harming others is not limited to those immediately present. In this context, spatial location is considered morally irrelevant. In deciding to whom the positive obligation to help in an emergency situation applies, it seems arbitrary suddenly to bring in spatial location as a morally relevant factor.

It may seem that neo-Kantian views such as contractualism can offer a plausible way of retaining a conception of the moral point of view as being fully impartial, and so avoiding setting arbitrary limits to the sphere of concern, while nevertheless avoiding utilitarianism's extreme demandingness. Since what the agent ought to do on each occasion is determined not by what will promote the most well-being or the least suffering, impartially considered, but by what is allowed and disallowed by the set of general moral principles to which everyone could reasonably agree, it follows that, if everyone has reason to reject highly demanding principles, the agent's moral obligations will avoid being extremely demanding. Moreover, the justification for the rejection of the demanding principles is impartial, since it is based not on anyone's bias in favor of her own interests, but on reasons that everyone in the position of an agent has. As I shall now argue, however, plausible versions of contractualism are in fact in the current state of the world no less demanding than utilitarianism.

D. The demandingness of contractualist obligations to help those in need. I shall focus on T. M. Scanlon's account of contractualism as he presents it in *What We Owe To Each Other,*[7] since I take his formulation to be the most fully developed and powerful version of contractualism as a theory governing individual moral obligations. But the structure of my argument is applicable to any plausible version of contractualism. According to Scanlon's contractualist formula, an act is wrong if it is disallowed by principles of conduct that no one could reasonably reject. The two most central aspects of his specification of reasonable rejection are, first, that the only reasons for and against a principle which are eligible for judging whether or not it can be reasonably rejected are "various *individuals'* reasons for objecting to that principle and alternatives to it" (*ibid.*, p. 229) and, second, that what counts as reasonable rejection depends on the comparative strength of individuals' reasons for and against the principle.

The demandingness of the contractualist principle governing obligations to help those in need principally depends on whether the specification of principles takes into account circumstances of partial compliance. The problem with ignoring circumstances of partial compliance in the specification of the principles is that this sacrifices what is particularly compelling about contractualism, namely, the idea that acceptable moral principles must be justifiable to each person who is burdened by them. Ignoring the circumstance of partial compliance involves ignoring the burdens faced by some actual persons as a result of the partial compliance.

---

[7] Cambridge: Harvard, 1998.

This can be shown by considering a modification of Scanlon's example of two swimmers struggling for the one remaining life jacket. Let us suppose that this time there are several swimmers in the water about to drown. Although there is a shortage of lifeboats, there is a surplus of life jackets stored in the lifeboats, which can keep those in the water alive until more rescue boats arrive. If each lifeboat distributes twenty life jackets, then all those in the water will receive one. A fair principle, which no one can reasonably reject, is that each lifeboat distribute twenty life jackets. Suppose, however, some of the lifeboats leave the scene without distributing any life jackets. If the remaining lifeboats were each to distribute only twenty life jackets, even though they had more left over, it is hard to see how this could be justifiable to those left to drown. Scanlon's contractualism is therefore most plausibly taken as holding that, in assessing which principle is justifiable to others, circumstances of partial compliance need to be taken into consideration.

Scanlon rejects the idea that there is a threshold level of cost, "such that it is reasonable to reject any principle that would lead to one's suffering a cost that great" (*ibid.*, p. 196). This is because of the stipulation that reasonable rejection depends on the *comparative* strength of individuals' reasons for and against a principle. In the case of the principle governing the obligation to help those in need, it will depend on the comparative strength of the agent's objection to a stringent principle of aid and of any other individual's objection to a less stringent principle.

In the current state of the world, there are constantly individuals in need whom a relatively affluent agent is in a position to help. This means that, if the agent complies with all but the most stringent possible principle of aid, there will be individuals in need whom she could help if she made a more substantial sacrifice. Scanlon describes those in need as "in dire straits: their lives are immediately threatened, for example, or they are starving, or in great pain, or living in conditions of bare subsistence" (*ibid.*, p. 244). Clearly, the sacrifice imposed on the agent by a principle of aid to those in need will have to be extreme before it balances the cost faced by individuals in dire straits who are not helped. The consideration of the comparative strength of the burdens faced by various individuals, then, will lead to principles of aid that cannot be reasonably rejected but which impose demands on the agent which are just as extensive as those imposed by utilitarianism.

Unlike utilitarianism, however, contractualism will take into consideration another objection which a relatively affluent agent can propose to a highly stringent principle of aid, which has considerable initial intuitive force. This is the complaint that it is unfair for the agent to face a huge burden of giving help because of others' failure to give any help at all. Whether this complaint grounds reasonable rejection of a highly stringent principle of aid depends on whether it outweighs the objection to a less stringent principle from those in dire straits whose vital interests could be saved if the agent were to give more than is required by the less stringent principle.

In the modified life-jacket example I described earlier, those in the remaining lifeboats might object to a principle requiring them to give more than twenty life jackets on the grounds that it is unfair for them to have to pick up the tab for others' failure to do their duty. Their objection would clearly be outweighed, however, by the complaint of those left to drown if those in the remaining lifeboats left after distributing only twenty life jackets. This indicates that in situations in which others' vital interests are at stake, agents' complaint that a principle requiring them to help is unfair is likely to be outweighed by the complaint of those who will die without the agents' help.

In that example, the extra cost of the obligation to help which is imposed by others' negligence is trivial. Let us therefore modify the example and suppose that the weather conditions are worsening, so that the conscientious agents who remain on the scene distributing the extra life jackets face a significant risk of capsizing, and are unsure whether the rescue boats will arrive in time to save all those in the water.

Even in this situation, in which others' negligence imposes a burden on the conscientious agents as great as a significant risk of death, their objection to a principle requiring them to distribute their surplus life jackets is outweighed by the objection to a less demanding principle of aid on the part of those who are certain to drown unless they are given a life jacket. This is because the behavior of those who left the scene without distributing any life jackets is at least as unfair to those left in the water as to the conscientious agents who are left with a highly onerous obligation to help those in need; those in the water are in an even worse position than those in the remaining lifeboats and equally share no responsibility for the situation. Since those in the water can make at least as strong a complaint about the unfairness of their predicament as the agents in the remaining lifeboats, the complaint of unfairness made by each member of the two groups balance one another. That leaves as the most salient remaining ground for determining the acceptable principle the comparative costs it imposes on the agents in the remaining lifeboats and on those in the water who will drown unless they are helped; and the cost to those in the water of certain death outweighs the cost of the risk of death to those staying behind.

Similarly, many of those in need in the current state of the world are in a desperate situation for which they are not responsible. Their complaint about the unfairness of their predicament as the result of others' negligence therefore balances the complaint by conscientious agents giving help that it is unfair that others' negligence makes the obligation of giving aid extremely onerous. And as we have seen, the consideration of the comparative strength of the burdens faced by those in need and those giving help grounds principles of aid which cannot be reasonably rejected and which are just as demanding as utilitarian obligations to help those in need.

I conclude that in the current state of the world, any plausible moral theory has difficulty in showing how agents' impartial moral commitments and their personal commitments can be harmoniously integrated. I shall now argue that it is a strength rather than a weakness of utilitarianism that it explicitly acknowledges the current threat to agents' objective integrity.

E. The appropriateness of utilitarianism's acknowledgment that our objective integrity is currently compromised. It is in emergency situations that agents' personal and moral commitments fundamentally clash. The reason for this is that in such situations, there is a conflict between different persons' pursuit of their basic projects. The agent's pursuing important personal projects conflicts with her helping others' even more urgent projects (such as continued survival, or not seeing their children die). This leads to a fundamental and irresolvable conflict between her pursuit of her personal projects and of her moral commitments to help others.

An example of such an extreme situation is Nazi Germany. Many agents were in a situation in which their complying with strong moral obligations to rescue Jews, or even to avoid collaboration with the Nazi regime, would endanger themselves and their family. Failure to comply with the moral obligations must often have led them to feel morally compromised and have had a very negative impact on their sense of moral worth. But their sense of identity and of the meaningfulness of their lives could also be severely marred if their actions led to their family's being harmed. This situation must have caused considerable psychological conflict.

We might consider these extreme situations, in which agents' objective integrity is threatened, to be cases of moral bad luck. The threat to integrity in the current state of the world can be seen as arising from the fact that we have the moral bad luck to be living in a constant emergency situation, which results in an irresolvable conflict between fundamental moral and personal commitments.

Utilitarianism makes this conflict explicit, by emphasizing the moral seriousness of omissions. According to utilitarianism, when an agent fails to help another's urgent interests in order to pursue her own less serious interests, this should be seen as constituting a trade-off (albeit a passive one) between her own and the other person's interests. In the current emergency situation, in which relatively affluent agents are continually in a position to save others' vital interests, utilitarianism therefore holds that tragic trade-offs are constantly taking place between central aspects of persons' well-being. The irresolvable conflict between agents' personal and moral commitments is the result of these trade-offs.

Utilitarianism's acknowledgement that in the current state of the world agents' objective integrity is inevitably threatened is, I suggest, entirely appropriate. When agents' pursuit of their ground projects conflicts with their saving others' vital interests, this poses a problem for agents' leading integrated good lives which should not be viewed as resolvable. Those facing extreme poverty have a claim that their vital interests be met which cannot be ignored. This continual crisis imposes a constant and very strong moral pull on agents who are relatively affluent and are as a result able to offer help. Nevertheless, the relatively affluent agents also have a competing reason not to sacrifice to the point of undermining their personal ground projects. Agents merely reach a compromise between the two commitments. Only those who are extraordinarily altruistic will choose to live Anna's third life. It is appropriate, however, that the rest of us who fail to do so should feel to some extent morally compromised, by our awareness that we are pursuing our personal projects at the cost of our failing to do all we can to help others' vital interests, and that from the moral point of view every person's interests are of equal weight. The only solution to this conflict is the eradication of the extreme poverty, which is the most urgent goal of utilitarianism because of the extremity and scale of the suffering it causes.

I conclude that Williams's integrity objection on the second interpretation also fails to establish that utilitarianism is too demanding.

IV. THIRD INTERPRETATION: THE THREAT TO OBJECTIVE INTEGRITY IN ANY PRACTICALLY REALIZABLE STATE OF THE WORLD

My third interpretation of Williams's integrity objection is that the threat utilitarianism poses to agents' objective integrity is deeper than its extreme demandingness in the current state of the world. The objection, on this interpretation, is that the very structure of the theory is incompatible with agents' integrity. It is therefore not simply because of the current emergency situation that utilitarianism threatens agents' integrity. This interpretation is suggested by passages in which Williams claims that agents' acknowledging and complying with impartial utilitarian moral reasons is incompatible with their being sufficiently emotionally engaged in their own lives for ground projects (in particular deep personal relationships) to even be possible for them.

A full assessment of this charge involves discussing it in the light of a characterization of the most plausible version of utilitarianism, which is beyond the scope of this paper. In the remaining section, I shall simply offer a few suggestions as to why a plausible version of utilitarianism would not threaten agents' objective integrity, if the state of the world were different.

Williams argues that agents' emotional engagement in their personal projects and commitments will be undermined by their impartial utilitarian reflection, as the result of a dissociation between agents' emotional engagement in their personal projects and their impartial utilitarian reflection. He claims that this will in practice undermine their ability to have such personal projects. But, although, as I argued in the last section, agents' pursuit of their personal and of their utilitarian moral commitments may regrettably conflict, there is no dissociation between them.

I shall consider, first, agents' prudential personal projects. Agents' special concern for these projects arises from their personal point of view, from which agents have a deep concern for how their own lives go. The impartial moral point of view differs from the personal point of view simply by the addition of the claim that every agent's interests have equal weight. The impartial perspective of utilitarianism does not transcend and cast off the value agents' lives have for those agents; rather, it is a widening of viewpoint, to include the value of each agent's life, assigning none priority. All the value from the agent's personal perspective is captured by the impartial perspective. When the agent adopts the impartial perspective, her life and ground projects have the same enormous value as they do from her personal perspective, but she is keenly aware that everyone else's life is of equal value. She is therefore not dissociated from her concern for her personal projects when she adopts the impartial moral point of view.

I shall now consider agents' commitments to particular individuals, with whom they have an intimate personal relationship. These commitments, I suggest, far from being dissociated from agents' impartial moral commitments, actually inform them. It is only in the context of a loving personal relationship that our concern for a person is sufficiently intense to mirror the moral importance of how another person's life goes.[8] The impartial moral point of view, again, simply holds that everyone else's life shares the same importance. Since it is psychologically impossible for us to have this intensity of concern for the whole of humankind, adopting the moral point of view involves our extrapolating from our concern for the lives of those we love (and for our own lives). It does not involve a diffusal of our love for particular individuals, because the resulting concern would be too weak to reflect fully the actual moral importance of others' lives. The intense love we feel for particular individuals is therefore not a cognitive error, but rather forms the basis of our grasp of the impartial moral point of view.

Moreover, the cognitive role of an agent's concern for particular individuals is not limited to its informing the agent's grasp of the moral point of view; it also enables those individuals to gain a full awareness of the importance of their own lives. This is most striking in the case of parents' love for their children. It is through their parents' love that children come to have a corresponding grasp of and concern for the value of their own flourishing.[9]

Finally, it is only in the context of intense loving relationships that what Crisp[10] calls the "intricacies of well-being" are promoted. In summary, it is through deep personal relationships that the value of persons' welfare is both fully grasped and promoted. This means that the utilitarian moral point of view will see as a tragic loss any situation in which, for example, children are deprived of that intense love, since it means that their well-being is not grasped and realized in such a way as to correspond to its full moral importance.

Utilitarianism will therefore assign considerable weight to agents' personal projects and relationships. They will be outweighed only by others' very serious interests. Williams's objection that the utilitarian agent must regard his ground projects as merely "one set of satisfactions" among others somewhat trivializes the seriousness with which they are viewed in the utilitarian calculus. The calculation assigns them significant moral priority over the short-term pleasures connoted by "satisfactions."

The phrase 'set of satisfactions' in addition suggests a conception of utility as an impersonal and coldly mathematical notion, transcending the ends of individuals. This overlooks the fact that utilitarian moral demands are for the sake of the interests and ground projects of particular individuals. The stringency of these demands in the current situation arises from the fact that extreme poverty threatens many persons' most basic interests and ground projects, such as not seeing their children die from preventable disease.

---

[8] The version of utilitarianism I take to be most plausible views concern for persons' welfare as derivative of concern for the persons themselves.

[9] Numerous empirical studies have shown that children who are deprived of such love are liable to develop tendencies to low self-esteem, self-abuse, and depression.

[10] "Utilitarianism and the Life of Virtue," *Philosophical Quarterly*, XLII (1992): 139-60, here p. 158.

In situations of trade-off between different persons' central interests, utilitarianism may prescribe that agents strive to detach themselves from their relationships of partiality sufficiently to be able to help others' urgent interests. In the current emergency situation, utilitarianism holds that parents' pursuit of their commitment to their children can conflict with their moral obligation to reduce the number of other children who die as the result of lack of nutrition and basic medical treatment. In a world in which there were not widespread easily preventable suffering, however, situations in which the agent was in a position to help others' basic physical interests would be rare. This would be brought about by the kind of reforms Thomas Nagel advocates in *Equality and Partiality*.[11] These reforms would mean that persons' physical needs would generally be taken care of at the political level, which would remove the continual onus on individual agents to help such interests.

In such circumstances, emergency situations and the corresponding obligation to give help would arise occasionally; but they would not occur on such a scale as to interfere seriously with agents' engagement in their intimate relationships. Impartial obligations to help others' less urgent interests would not be so stringent as to conflict seriously with agents' pursuit of their close personal relationships and other ground projects, given the importance of these personal commitments from the utilitarian moral point of view. Agents would therefore generally not have to choose between pursuing their ground projects and complying with their impartial moral obligations, but could integrate both into a personally fulfilling and morally good life. Their reflection on and compliance with their impartial moral obligations would enhance their sense of the meaningfulness of their lives and give a firm and genuine grounding to it.

In the current constant emergency situation, we tend to engage in what Peter Unger[12] calls "futility thinking." We feel swamped by the vastness of the number of those whose vital interests we fail to help, rather than focus on what we are able to accomplish positively (as we do in the case of saving a drowning child from a pond). This leads to a feeling of impotence and failure. Moreover, we tend to find the scale of the problem psychologically overwhelming, and to take refuge in psychological distancing mechanisms. This tends to diminish our awareness of the full force of the moral reasons to give help. Since objective integrity requires a grasp of the reality of our moral obligations and a genuinely grounded sense of accomplishment, psychological distancing and futility thinking both threaten it. These two psychological responses would be far less common if emergencies were to arise occasionally and on a small scale, and we were able to give help whenever they did.

---

[11] New York: Oxford, 1991.

[12] *Living High and Letting Die* (New York: Oxford, 1996), pp. 75-82.

In the present situation, if we address the urgency of the crisis and meet some of our obligations to give help, we may feel to some extent morally compromised because of our awareness that, short of leading a life like Anna's third one, there are so many urgent moral obligations with which we are failing to comply. Nevertheless, we shall be able to achieve a degree of objective integrity, through honest appraisal of our obligations and meeting some of them. By contrast, if we do not face the reality of the urgency of our obligations, this will completely preclude our having objective integrity.

The ultimate goal toward which we should be striving is to eliminate the tragic trade-offs that currently to a greater or lesser degree compromise our objective integrity.

ELIZABETH ASHFORD

Wolfson College/University of Oxford

SURPRISE, SELF-KNOWLEDGE, AND COMMONALITY*

Here is a familiar puzzle. A teacher announces on Monday that there will be a surprise exam on either Wednesday or Friday. Her students reason as follows. Say that the teacher's announcement is true. Then, if the exam were on Friday, we would know by Thursday, and it wouldn't be a surprise. Therefore it won't be on Friday. This means it must be on Wednesday, and since we know that, it can't then surprise us. There can't be a surprise exam on either of these days. The teacher's announcement is false. It contradicts itself.

The teacher gives the exam on Friday and everyone is surprised. Where did the students go wrong? This has been much discussed, and I want to discuss it once more. I want then to extend my discussion to some larger issues.

I

Let me bring it down to just a single student's problem. The announcement can be put as three propositions. There will be an exam on either Wednesday or Friday:

(1) $W \vee F$

And it will come as a surprise to this student. If, that is, it will be on Wednesday, he will not, on Tuesday, believe it will be on Wednesday. And if it will be on Friday, he will not believe, on Thursday, that it will be on Friday:

(2) $W \supset \neg B_t W$
(3) $F \supset \neg B_{th} F$[1]

Let us suppose also that both

(4) $\neg(W \cdot F)$ and
(5) $\neg W \supset B_{th} \neg W$

Assume now it will be on Friday. This begins a conditional proof:

(6) F
(7) $\neg B_{th} F$ (6) and (3)

---

*An earlier version of this paper was one of a number of papers written for Peter Gärdenfors on his fiftieth birthday. These papers all are posted at www.lucs.lu.se/spinning.

[1] Imagine a second subscript to the "B"s identifying the believer—in this first case, the student.

0022-362X/00/9708/440–53

(8) $\neg W$ (6) and (4)
(9) $B_{th}\neg W$ (8) and (5)

We want to proceed to $B_{th}F$, but (9) and (1) do not warrant that, for perhaps the student won't believe (1) on Thursday. We need to add the supposition that

(10) $B_{th}(1)$

This lets us move the conditional proof to

(11) $B_{th}F$ (10) and (9)

which, by reductio, gives us:

(12) $\neg F$ (11) and (7)

We go on to

(13) W (12) and (1)
(14) $\neg B_t W$ (13) and (2)

These two lines say that the exam will be on Wednesday and will surprise the student. No contradiction here.

But suppose also that, on Tuesday, the student believes all our suppositions above:

(15) $B_t(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10)$

We then get

(16) $B_t W$ (15) and (1) through (13)

Here we do have a contradiction: (16) contradicts (14). But it does not follow that the teacher's announcement was false—this is where the students went wrong. What follows (by reductio) is that either that *or something else used in the argument* is false. That is, what follows is:

(17) $\neg(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10 \cdot 15)$

There is a tacit assumption here. We are supposing that the student is deductively *thorough*, that he believes the deductive consequences of the conjunction of all he believes, that he is in this sense thorough every day of this week.[2] Suppose now also that his beliefs are *stable*, or at least stable this week, barring new relevant information—that he is belief *retentive*, that he gives up no beliefs unless he gets such information. This allows us to simplify.

---

[2] We need this in moving from (9) and (10) to (11) and also in moving from (15) to (16).

Let *m* be Monday, *today*. Then (15) can be replaced by

(15′) $B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 10)$ and also by
(15″) $B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)$

for (15″) now implies (10), which thus becomes redundant. So (17) reduces to

(17′) $\neg(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 15'')$

which is equivalent to

(18) $(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5) \supset \neg B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)$ and to
(19) $B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5) \supset \neg(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)$

Nothing is wrong with the teacher's announcement, nor indeed with $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$; certainly the conjunction is not contradictory. But there clearly *is* something wrong with $B_m(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)$. Let us describe someone who is both deductively thorough and belief retentive (during this period) as *disciplined*. We have seen that, if $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$ is true, a disciplined student does not believe it (this is (18)). And also—this says the same—that if he believes it, it is not true (19). That is, he *cannot* believe it, not in good logical conscience: there is no possible world in which he believes it and it is true.

Here is a simpler scenario. The teacher says there will be a surprise exam on Wednesday.[3] She tells the student that

(20) W and
(21) $\neg B_t W$

No problem with $20 \cdot 21$; perhaps the exam *will* be on Wednesday and will surprise the student. But if

(22) $B_m(20 \cdot 21)$, then
(23) $B_t W$ (22)

This contradicts (21), so (by reductio) it follows that

(24) $\neg(20 \cdot 21 \cdot 22)$, and so too that
(25) $(20 \cdot 21) \supset \neg B_m(20 \cdot 21)$ and
(26) $B_m(20 \cdot 21) \supset \neg(20 \cdot 21)$

---

[3] This simpler case is in W. V. Quine, "On a So-Called Paradox," *Mind*, LXII (1953): 65-77, as is the line I take just above. Even simpler is G. E. Moore's case of the person who says it is raining but that he does not (now!) believe it.

Nothing is wrong with the announcement here either. Still, the student should not have believed it. As with 1 · 2 · 3 · 4 · 5, 20 · 21 is not contradictory, but the student could properly believe it only if he were not disciplined.[4] For he would be believing a proposition that can be shown to be false if he believes it and is disciplined. Jaakko Hintikka[5] calls such propositions *doxastically indefensible* for this person. I shall call them *incredible* for him, or not properly *believable* by him. He cannot *properly* believe such propositions because he could defend his believing them only by admitting to a lack of discipline. Another way of putting it: if he believed such a proposition and also believed he believed it, he would have to believe a contradiction (like 14 · 16 and 21 · 23).

I shall stretch the concept of discipline to cover noncontradiction, the *non*belief of *x*-and-not-*x*; this makes what is logically false incredible for all. Still, we have been speaking of incredible propositions that may in fact be true—the teacher's announcement *is* true—and these are incredible for some people only. Thus 20 · 21 is incredible for the student but is not incredible *for me*, there being nothing in that announcement about *my* being surprised. *I* can properly believe it, the student involved cannot. Getting beyond our scenarios: I can predict that this or that will surprise someone else, but no one can say about himself (can properly believe) that *x* will surprise him.

Let me touch on two small points. Is the above about *surprise*? I held that *x* would surprise the student if he did not believe it beforehand. Perhaps this puts it badly; perhaps a person is only surprised if, beforehand, he believed *not-x*. No problem in that for us here. Suppose that the teacher meant that her students would be surprised in the fuller sense. Strengthening her announcement to make that explicit cannot weaken its force, so our conclusion would still stand. What the teacher said would still be incredible for the student.[6]

Also, the title mentions self-*knowledge*, but the above has brought out only that we cannot properly *believe* certain propositions. Still, if we cannot believe them, we cannot know them either. We can't have full self-knowledge because we can't have full self-belief. We aren't fully knowable to ourselves because we aren't fully credible to ourselves.

---

[4] In the second situation, we are assuming discipline in the move from (22) to (23).

[5] *Knowledge and Belief* (Ithaca: Cornell, 1962). For Hintikka, being disciplined does not extend to being belief retentive.

[6] The proof would need some extra moves from $B\neg x$ to $\neg Bx$, but such moves are warranted by thoroughness plus noncontradiction.

II

The surprise exam conundrum is of no interest in itself. That is why I discussed it. I counted on no one's really caring how things came out in that situation to let the reader be receptive to my analysis of it. I want now to apply that analysis to a different situation.

My wife knows me very well. Still, she sometimes gets me wrong. Say that I am agonizing whether to do x or y, and she tells me the outcome is clear: I will choose to do x.

I now think about this. She says that, on Wednesday, I will choose x. This x is indeed an option for me. That means it is (*a*) an action I neither yet think I will take nor that I won't. It is also (*b*) an action I think I would take if I chose it. So, if I accepted my wife's prediction that I will choose x, I would (by *b*) now think I would take it. But then (by *a*) x would not be an option, and neither would y be an option. There would be nothing for me to choose. So—aha!—she is wrong. She has contradicted herself! And I go on agonizing until, on Wednesday, I choose to do x. Her prediction turns out to be right; it was *not* contradictory. Where did my reasoning fail?

I will stand by (a) and (b). A set of options composes an *issue*, and an issue is a situation it makes some sense to agonize over. There is no sense in agonizing where we know what we will do—for instance, in sweating out the question of whether or not to have dinner tonight. And it makes no sense to agonize where we think we are weak, where we doubt we would follow through. There is no point in Romeo's asking whether he ought to leave Juliet: he doubts he would leave her even if he chose to, so leaving her is not an option for him. I take this to mean that (a) and (b) are basic to optionality.[7]

Where did my reasoning fail? My wife predicted I would, on Wednesday, choose x—she predicted

(27) $C_w x$

Since what I will choose will have been an option,[8]

(28) x will on Tuesday be an option for me (27), and
(29) $\neg B_t x$ (28), by (a), and
(30) $B_t(C_w x \supset x)$ (28), by (b)

This alone is not troublesome. But if I *believe* the prediction, believe it now, on Monday—if, that is,

---

[7] For more on optionality, see my *Making Choices* (New York: Cambridge, 1997), pp. 8-11.

[8] Strictly, it will have been an option *as I understood it*; see my *Making Choices*, pp. 11-20. I shall not press this refinement here, but we need it for getting around some objections.

(31) $B_m(27)$, then
(32) $B_t C_w x$ (31) and
(33) $B_t x$ (32) and (30)

Since (33) contradicts (29), something has to go. But nothing is wrong with (27)—my thinking failed in thinking that there was. I should have concluded that

(34) $\neg(27 \cdot 31)$, and so too that
(35) $(27) \supset \neg B_m(27)$ and
(36) $B_m(27) \supset \neg(27)$

Suppose I am fully disciplined.[9] Then, if (27) is true, I do not believe it (this is (35)). And if I believe it, it is not true (this is (36)). My wife may be right to believe (27), but (27) is not credible *for me*, for *my* believing it would make it false. Generalizing again: everyone can say about other people that they will choose certain options they have, but no one can say about himself (can properly believe) that he will choose x.[10]

We cannot predict our own choices any more than we can predict our surprises. But this is not to be wondered at, for choices are a sort of surprises (in my weak sense of the word). In choosing, we surprise ourselves; we come in the end to do what, before, we did not think we would. And there is more of the same. We cannot predict our *learnings*. No one can properly believe he will learn (or conclude, or discover, or realize) x—a point stressed long ago by Karl Popper.[11] For suppose you *will* learn x, say on this coming Wednesday. You can't learn what you already believe, so you won't yet believe x that morning. Still, you surely will then believe that whatever you learn is true. No contradiction here; you may in fact learn x. But if you *believe* you will learn it, the argument continues as in (32) to (36). Thus your learning x on Wednesday is now incredible for you. Since the truth of x will surprise you, you cannot predict your learning it.

We can take this beyond *prediction*, beyond our foreknowledge (our fore*belief*) of our own choices and learnings and the like. For if k is not properly believable and *h* implies *k*, then *h* is not believable either. Say that *x*, *y*, and *z* are your options, c_1, c_2, c_3, c_4... are their consequences in different states of nature, *P* is your probability distribution and *U* your utility distribution, and that you are rational (the last clause is more about discipline). Let all this be *h*. Say that it follows from *h* that *k*: that you will choose *x*. Since *k* is not believable for you, neither is *h*. But *h* is not a proposition about some choice you will make, nor does it speak of anything else that will come to surprise you.

---

[9] We need deductive thoroughness to move from (30) and (32) to (33). We need belief retentiveness to move from (31) to (32).

[10] A person cannot predict his choices only *in the terms in which he sees his options*; I note this in my "Self-knowledge, Uncertainty, and Choice," *British Journal for the Philosophy of Science*, XXX (1979): 235-52; reprinted in Peter Gärdenfors and Nils-Eric Sahlin, eds., *Decision, Probability, and Utility* (New York: Cambridge, 1988), pp. 270-86. Romeo may be able to predict that he will choose as Juliet would have chosen or as she now wants him to—provided he cannot infer from that which of his options *as he sees them* he will choose.

[11] "Indeterminism in Quantum Physics and in Classical Physics," *British Journal for the Philosophy of Science*, I (1950): 117-33 and 173-95.

We can take it further still, beyond any connection with choosings or learnings. Say that you have forgotten *x*. You cannot forget what is false, and cannot forget something and also believe it, so *x* is true and you don't believe it. You cannot properly believe this conjunction (that *x* is true and you don't believe it)—that is like 20·21. And since the conjunction follows from your having forgotten *x*, your having forgotten *x* is not believable for you either.[12]

Again, these are limits to *self*-belief and self-knowledge. *Other*-knowledge is not so restricted. Jill can know what Jack will choose. Still, she cannot convey that knowledge she has about Jack to Jack—such knowledge is out of bounds for him. This reverses the usual thesis on people's mental privacy. The thesis is that there is much about us that we can know but no one else can, unless we choose to tell them (for instance, that we now have an itch). The point here is the opposite, that other people can know things about us that we cannot know, even if these others tell us. Privileged access goes in both directions.

The point is not wholly new. Suppose that

(37) Jack believes nothing

And suppose that Jack believes this, that

(38) $B(37)$

Then it is false that Jack believes nothing

(39) $\neg(37)$ (38)

This contradicts (37), so it follows that

(40) $\neg(37 \cdot 38)$, and also that
(41) $(37) \supset \neg B(37)$ and
(42) $B(37) \supset \neg(37)$

---

[12] The logic of forgetting is discussed in Hans K. Hvide, "Bounds to Memory Loss," *Theory and Decision*, XLVI (1999): 1-21.

Jack may in fact be a total skeptic: (37) may be true. Jill can believe that Jack is a skeptic. But he himself cannot properly believe it, for he cannot rightly believe what must be false if he believes it. The moral of this Cretan-like case (the Cretan says that he always *lies*) has to do with self-reference. The argument shows that certain propositions referring to a specified person p cannot properly be believed by that person. So also with our other arguments: these too involve some p's believing some proposition referring to p. They too reveal the impropriety of certain beliefs that are self-referential. The only novelty in the new cases is that *these* self-referential beliefs (seldom under any suspicion) are revealed as improper.[13]

III

I can hear a question: "OK, we can't predict our choices, but we can predict our *actions*. We can't predict we will choose x, but we can predict we will *do* it, that we will act out x: believing we will do x isn't improper. True, if we believe this (and are belief retentive), we can't go on to *choose* x. (Since $C_t x \supset \neg B_m x$, $B_m x \supset \neg C_t x$.) We can only think we will x, *properly* think this, if we won't choose x. But why should that concern us?"

First, because it has to affect what we think of ourselves as thinkers. The point is that we sometimes choose, and that we cannot then know beforehand *what* we are going to choose, can't even know what we will *do*—or even properly *believe* we will choose it or even that we will do it. Others may know this, but we ourselves can't. To that extent, our knowledge is bounded, and bounded not by our mental limitations but by our self-discipline.

This may leave you unconcerned, so here is a second point. We can indeed deny ourselves the beliefs that I say are improper. We lose nothing of any importance if we avoid such beliefs. Still, we sometimes ascribe such beliefs, if not to ourselves then to others. Sometimes we even endorse ideas that oblige us to do that. We trip ourselves up where we do, so it is well to be cautioned against it.

Let Jack and Jill be in a prisoner's dilemma. They can either cooperate or not—not cooperating is *defecting*. For each, defection is the dominant option and the probabilities of the other's options are independent of his (her) own. Both Jack and Jill are rational, so they both will defect. And since they both prefer the outcome of joint cooperation to that of joint defection, they will both be sorry.

---

[13] Some improper beliefs are *not* self-referential. Think of the antimentalist theorist who believes that there are no beliefs.

Where they think they will not meet again, there is no way out for them. But say that they think they will meet again, that their present interaction is only the first of many just like it. Here it may seem that the prospect each faces of having to live with the other's resentment ought to deter defection. Still, it often is argued that, where the number of rounds is finite and known to both of the agents, if they are rational they both will defect from the start to the finish.

The argument is this. Suppose that the number of rounds is known by both Jack and Jill to be 100, and that they both are rational. In the 100th, each will know that there will be no further meetings, no need to guard against reprisals, so they both will defect. Let Jack think that Jill is rational (that she will be rational throughout). He will then think in the 99th round that Jill will defect in the 100th whatever *he* does in the 99th, that his cooperating in the 99th round would not be rewarded by Jill in the 100th. He will therefore defect in the 99th, and Jill, thinking likewise about Jack, will too. The same in round 98: each expecting the other to defect in the round that follows, each will here defect, though we must now also assume that Jack thinks that Jill thinks *him* rational—and must also assume that Jill thinks he thinks *her* rational. So we move stepwise back to round 1 (both agents defecting all the way), though with a heavier load of assumptions at each preceding stage.

This backward induction rests on assumptions about these people's beliefs about each other. Sometimes these are put as follows: both Jack and Jill believe that they both are rational (that they will be rational throughout), and that, in each round, they each face a problem of the prisoner's dilemma sort, and that there will be 100 rounds. Both also believe that they both believe this, that both believe that both believe it, that both believe that both believe that both believe it, and so on. All this together is called the *common belief assumption* (CB), strictly: the assumption of the commonality of their beliefs in their both being rational and about the structure of their interaction.

We can now see that CB is too strong.[14] CB implies that Jack believes that he is rational (and will be throughout) and that in the last, 100th round he will be in a prisoner's dilemma. From that, it follows that he will there choose to defect, for in a dilemma only defection is rational. Being disciplined, Jack *believes* he will choose this. He believes that *now*. But then he *cannot* choose it—adapting

---

[14] It has been noted that the 'and so on' goes too far: only 99 iterations of "both believe that..." are needed in a 100-round case. My point here is different.

(36): $B_1(C_{100}x) \supset \neg C_{100}x$. So also for every previous round; he can never choose to defect. By the same argument, neither can Jill.

This undermines our scenario. For the basic dilemma premise is that Jack and Jill have options, that they each must choose. Given CB (and their both being disciplined), this basic premise cannot stand. Each of them knowing what they will do, there is no issue for either of them. Neither has any choice to make: they are in no dilemma. If we want to speak of dilemmas, we must give up CB. We cannot use it in discussing dilemmas.

The remedy is clear: we need a suitably weaker thesis, a thesis that, unlike CB, ascribes (imposes) no improper beliefs. Say that *s* is a set of propositions about the parties involved being rational and about the structure of their interaction. Suppose that each person believes every item in *s* that he can *properly* believe (in our special sense of *propriety*), that each believes this of all the others, that each believes that all the others believe it of all the others, and so on. Call this the *reciprocal belief assumption* (RB), strictly: the assumption of the reciprocity of their beliefs regarding the items in *s*. (CB with regard to these items implies RB with regard to them, but not vice versa.)

Where *s* is as above, RB implies that Jack believes that Jill is rational (plus other facts about her). It implies that Jill believes that Jack is rational (plus other facts about *him*). It implies that Jack believes that Jill believes that Jack is rational, that Jill believes that Jack believes that Jill is rational, and so on. It does *not* imply that either believes that he (she) himself (herself) is rational, or that the other believes himself (herself) to be rational, and so on. No improprieties here, or ascriptions of them to others (or any ascriptions of any beliefs whatever to oneself or of any self-ascriptions of any beliefs to others...). But note that this suffices for the backward induction, for the outside observer's argument to the joint-defection prediction. The observer's backward induction does not call for CB. It needs only the weaker RB—and the *observer/predictor's* beliefs about both parties being rational, and so on.[15]

This thinking can be extended. CB is often said to be essential to the theory of games, or at least to justifications of equilibrium solutions. If that were right, it would mean trouble, for it would mean that game solutions are justifiable only where there are no solutions, where the parties have no options, where there is no game to play. I suggest that it is wrong, that CB is more than game theory needs, that the jobs it is asked to do can be done by RB. Ken Binmore and Adam Brandenburger[16] say that "any equilibrium notion that incorporates some measure of self-prophesying necessarily entails common [belief] requirements" (*ibid.*, p. 106). I am saying that equilibrium analysis calls just for *other*-prophesying, that a player is not also (can't be!) the observer/predictor of the game. The moral here is the same as above: the logic of proper, defensible belief denies us nothing we need, but neglecting that logic invites an undermining confusion.

---

[15] It may be that even RB is too strong; this is implicit in the critique of CB in Philip Pettit and Robert Sugden, "The Backward Induction Argument," this JOURNAL, LXXXVI, 4 (April 1989): 169-82; and Cristina Bicchieri, *Rationality and Coordination* (New York: Cambridge, 1993). If RB is too strong, the backward induction fails, but not now because the parties have no issues to resolve.

The moral is not one for game theory only, for CB is sometimes endorsed in other contexts too. John Rawls[17] proposes what he calls a *publicity* condition. Initially, he speaks of that in connection with a system of rules: "A person taking part in an institution knows what the rules demand of him and of the others. He also knows that the others know this and that they know that he knows this, and so on" (*ibid.*, p. 56). Still, he later goes beyond rules: "for the most part, I shall suppose that the parties [behind the veil of ignorance] possess all general information" (*ibid.*, p. 142).[18] This means that everyone knows that everyone has *x*, *y*, and *z* as options, options distributing primary goods in certain specified ways, that everone prefers more such goods to less, and that everyone is a maximinner. If everyone had this information, everyone could predict his own choice, which is not logically possible (it is not possible for *any*one). Better: there would be no choices to make. There would be too much knowledge behind the veil of ignorance to allow for any choosing.

No problem here for Rawls, for he retreats from his supposition, though only on the grounds of the "complexity" of some of the information behind the veil. I am arguing there are *logical* grounds for backing off from publicity, from CB-publicity, that if Rawls assumed it, he would undermine his own theory. Fortunately, he does not need it, he does not need to suppose that everyone behind the veil is that well informed. (He does not even need the weaker, RB-sort of publicity.)

---

[16] "Common Knowledge and Game Theory," in Binmore, *Essays on the Foundations of Game Theory* (Cambridge: Blackwell, 1990), pp. 000-000. Binmore and Brandenburger are speaking not of common belief but of common *knowledge*, and there reciprocity implies commonality. Since what is known must be true, if Jack knows that Jill *knows* he is rational, Jack himself knows he is rational. But game theory needs only to talk of beliefs.

[17] *A Theory of Justice* (Cambridge: Harvard, 1971).

[18] Their all having that information is then itself a general datum, and they have to believe *that*, and so on.

Once more, the moral is this. There is no purpose for which we ever need to hold improper beliefs, and none for which we must suppose that other people hold any. But we mess up our thinking badly if we ascribe such beliefs to people, whether ourselves or others.[19]

IV

Let me briefly comment on some ramifications. We have spoken of the skeptic who believes nothing whatever. He cannot believe this about himself—can't *properly* believe it—for if he believed it, it wouldn't be true. Consider now his sister, who has joined a cult that requires its members not to *want* anything. Can she want to comply with that? If she wanted not to want anything, her wanting this would defeat what she wants. I will say she can't *properly* want it: if it is true she wants nothing, she *doesn't* want it (for she wants *that*), and if she wants it, it isn't true. Another way of saying this: if she did want it and wanted to want it, she would be involved in a contradiction.

Take also the teacher who announces to her students that there will be a test on Wednesday and that its being on that day will be a disappointment to them. This means there will be a test on Wednesday and that the students will want beforehand (say, on Tuesday) that it *not* be on Wednesday. Can the students *want* this to be true? Can a person ever want to be disappointed by a specified *x*? I have argued that no one can properly believe he will be surprised by *x*. A parallel argument (with *wanting* in place of *believing*) shows that no one can properly want to be disappointed by *x*: if he will be disappointed by it, he does not want that disappointment, and if he wants that disappointment, he will not be disappointed, at least not by *x*.[20] Just as we cannot predict our surprises, so we cannot hope for our disappointments.

What about desire publicity? Are the desire analogues of CB and RB of any philosophical interest? The ethics of Immanuel Kant[21] comes to mind. He presented a formal criterion of the desires we could morally act on, one that abstracted from our actual situations and from those of all other people. His categorical imperative is this: "Act only on that maxim whereby thou canst at the same time will that it should become a universal law" (*ibid.*, p. 38), a *maxim* being a desire to take a certain action described as the agent *sees* or *understands* that action. The desires a person acts on *are* his maxims, so we might put it like this—in terms of respecting *desire* commonality—that no one ought ever to act on a desire that could not be held by all.[22] Still, this has no force: there is no desire that could not be held by all.

---

[19] Isaac Levi, in discussing my "Self-knowledge, Uncertainty and Choice," draws a different moral; see his *Hard Choices* (New York: Cambridge, 1986), chapter 4. He extends his analysis in *The Covenant of Reason* (New York: Cambridge, 1997), chapters 2, 4, and 5. I comment on Levi in my review of *The Covenant of Reason* in *Philosophy and Phenomenological Research* (forthcoming).

[20] The argument appeals to a desire extension of our concept of discipline. Deductive thoroughness carries over smoothly, and so does noncontradiction, but retentiveness needs some rethinking. (In default of what sort of new data must our desires be stable—what new data is *relevant* here?)

[21] *Fundamental Principles of the Metaphysic of Morals*, Thomas K. Abbott, trans. (New York: Liberal Arts, 1949).

But perhaps this sells Kant short. Perhaps he means that the maxim you act on must be one that could be held by all people who are logically disciplined. This would move us to desire *reciprocity*[23] and to this different version of the imperative, that no one ought ever to act on a desire that could not *properly* be held by all. If Jill couldn't properly want *x* true, Jack oughtn't to act on his wanting *x*. This again uses 'properly' in our special, logical sense—for Kant, morality looks only to logic. (The imperative itself is not logic, but, put in reciprocity terms, what it excludes depends just on logic.)

Here is an appealing corollary: never act on any desire to disappoint someone by *x*. We are led to this by the fact that whoever you want to disappoint cannot properly want to be disappointed by it. (The fact emerges in the proof of the impropriety of wanting that, so the "fact" is mere logic.) This does not say we may never disappoint any person. We cannot avoid disappointing; we do that whenever we do what anyone wants us not to be doing. It says we ought never to act *on the maxim* of disappointing someone (by *x*), that we may never make that our purpose, never may set our minds on just that. Are there other such corollaries? Not as things now stand with us, but let us move a step further.

We have taken people to be deductively thorough with regard to beliefs and also with regard to desires. Let us say that being disciplined also implies a thoroughness with regard to beliefs and desires together, though this has to be hedged a bit to keep it from being too strong. Let us here put it this way, that if a person is disciplined and the conjunction of all he believes is *h* and the conjunction of all he wants is *k*, then if *m* follows from *h*-and-*k* and no conjunct of *h* alone follows from *k*-and-*m*, he also wants *m*.[24] (If disciplined Jack wants the

---

[22] This refers only to commonality of degree 1, not to any CB-like iterations.

[23] To reciprocity of degree 1.

[24] The "no conjunct of *h*..." proviso is meant to avoid the implication that a disciplined person wants whatever he believes. The proviso in my *Understanding Action* (New York: Cambridge, 1991, p. 50) is too strong; this was noted by Christian Piller in his review of that book in *Erkenntnis*, XLI (1994): 127-33.

Democrats to win and he believes that their candidate is Jones, he wants Jones to win.) We now have another corollary, this one perhaps unexpected.

Say that Jill will, on Wednesday, be led to falsely believe not-*x*. And suppose she now, on Monday, wants (desires) this to happen then. Here we have

(43) $x \cdot B_w \neg x$
(44) $D_m(43)$

By desire thoroughness and retentiveness, we have

(45) $D_w x$ (44)

Jill will on Wednesday want x (45) and will believe $\neg x$ (43). Since everything follows from $x \cdot \neg x$, $\neg(43)$ follows. Belief-plus-desire thoroughness here gets us

(46) $D_w \neg(43)$ (45) and (43)

By desire retentiveness and thoroughness, we have

(47) $D_w(43 \cdot \neg 43)$ (46) and (44)

Since by desire noncontradiction, (47) is false, we have, by reductio,

(48) $\neg(43 \cdot 44)$ and also
(49) $(43) \supset \neg D_m(43)$ and
(50) $D_m(43) \supset \neg(43)$

These last lines say that Jill cannot properly want to be misled about *x*. Thus the categorical imperative implies that Jack ought not to try to mislead her: he ought not to lie to her. Kant would have said this is paydirt. He wanted to establish that lying is wrong—not that lying itself is wrong, but that an action taken on a lying maxim is wrong. Those who agree with that idea can take this last proof as a vindication of Kant. Those who *dis*agree can take it as a reductio of the categorical imperative, of the injunction to respect reciprocity, never to act on any desire that could not be properly held by all.

FREDERIC SCHICK

Rutgers University

COMMENTS AND CRITICISM

FODOR AND LEPORE ON MEANING SIMILARITY AND COMPOSITIONALITY*

In "All at Sea in Semantic Space: Churchland on Meaning Similarity," Jerry Fodor and Ernie Lepore[1] argue against the utility of a notion of meaning similarity as a replacement for meaning identity in theories of semantics. They argue that meaning similarity fails with respect to four desiderata, one of which is compositionality. This argument is not valid.

Compositionality is the determination of the meanings of complex expressions by the meanings of their constituent parts (plus the way those parts are combined syntactically). Thus, the meaning of 'brown dog' is determined (in part) by the meanings of 'brown' and 'dog'. Fodor and Lepore point out that compositionality depends on the context independence of constituent meanings—for example, the fact that "'brown' means the same in the environment '_dog' that it does in the environment '_cat'..." (386). Context independence is schematized in their principle (S2):

> S2. If *m* is part of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is part of the meaning of '*b*' (386).

Substituting the notion of meaning similarity, Fodor and Lepore derive (S3):

> S3. If *m* is similar to part of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is similar to part of the meaning of '*b*' (387).

Fodor and Lepore's argument that meaning similarity is not compositional is based on an attempt to show that (S3) is false. To do this, they give the following example:

> So, presumably, the meaning of 'cat' is similar to part of the meaning of 'The leopard is on the mat' (cats and leopards are both felines) and part of the meaning of 'The leopard is on the mat' is similar to part of the meaning of 'The explosion caused a lot of damage' (leopards and explosions are both dangerous). But the meaning of 'cat' is not similar to any part of the meaning of 'The explosion caused a lot of damage' (*ibid.*).

---

*I am grateful to Rich Hall and Larry Hauser for their comments on a draft of this note.

[1] This JOURNAL, XCVI, 8 (August 1999): 381-403.

0022-362X/00/9708/454–55

But this example is not an instantiation of (S3). If we let *m* in (S3) be the meaning of 'cat', and '*a*' be 'The leopard is on the mat', and '*b*' be 'The explosion caused a lot of damage', then '*a*' is not a constituent of '*b*'. There does not seem to be any other way to make the example instantiate (S3) either.

In fact, principle (S3) appears to be true for compositional expressions. If (S3) were not true, then there could be a meaning *m* which was similar to part of the meaning of an expression '*a*', but not similar to any part of the meaning of an expression '*b*' of which '*a*' is a constituent. Let *n* be the meaning part of '*a*' to which *m* is similar. If '*a*' is a constituent of '*b*' and '*b*' is compositional, then, by principle (S2), *n* must be part of the meaning of '*b*'. But there is no reason to suppose that *m* would no longer be similar to *n*.

The fact that principle (S3) holds for compositional expressions does not mean that meaning similarity *is* an appropriate notion on which to base a compositional semantics. Fodor and Lepore may still be right that it is not. One way to show that might be to show that (S3) is true for expressions whose meanings are intuitively not compositional. If that were the case, and if (S2) would not have been true for those expressions, that would indicate that (S3) is not a satisfactory replacement for (S2).

Consider the fact that many people who are correctly describable as having red hair actually have hair that is orange colored. 'Red', in 'red hair', means something similar but not identical to what it means elsewhere. Intuitively, the phrase 'red hair' is not completely compositional. Someone could know what 'red' means and know what 'hair' means, and still be surprised on first learning to what the phrase 'red hair' applies. The failure to predict meanings of combinations is the essence of idiomaticity (the lack of compositionality), and (S3) fails to rule it out.

On the other hand, some people may not find this argument against meaning similarity convincing. Compositionality is generally supposed to be the explanation for our ability to produce and understand novel utterances. It could be argued, however, that there is more idiomaticity in natural language than Fodor and Lepore believe to be the case. Possibly words commonly fail to be context independent (in Fodor and Lepore's sense), and we do a lot more guessing at the meanings of complex expressions than we do combining the unmodified meanings of the constituents. That issue goes beyond the scope of this note.

BARBARA ABBOTT

Michigan State University

COMMENTS AND CRITICISM

LEVELING THE PLAYING FIELD BETWEEN MIND AND MACHINE: A REPLY TO MCCALL

The temptation to think that somewhere in Gödel's incompleteness theorems there is a proof that the mind is not a formal system is seemingly irresistible.

But where? The natural place to look first is *G*, the Gödel sentence of PA (a formalization of Peano arithmetic). The proof of the first incompleteness theorem reveals that, assuming PA to be consistent, PA cannot prove *G*. But since *G* in some sense says of itself that it is not provable, we minded creatures can see that it is true. Thus, it is often claimed, we can recognize as true something that PA cannot prove, and so human thought cannot be modeled by PA. Similar reasoning applies to other formal systems, and so, it is alleged, human thought cannot be modeled by any formal system.

But it is an illusion to think that we unconditionally recognize *G* to be true; rather, our recognition of its truth depends on our assuming that PA is consistent. Without this assumption, there would be no basis for asserting the truth of *G*. Furthermore, if we allow PA to make use of that very assumption, then it, too, will be able to establish *G* (for $\mathrm{Con(PA)} \rightarrow G$ is derivable in PA). Our tacit reliance on the assumption of consistency is obscured in the case of the formal system PA, for its consistency just seems to go without saying. For other formal systems, consistency may be doubtful, and, if we are unwilling to assume the consistency of a system, then we will be unable to recognize the truth of its Gödel sentence. Thus, we have no evidence here that the mind cannot be represented by a formal system.

It is here that Storrs McCall[1] takes up the search for the Holy Grail of antimechanists. He suggests that instead of focusing on our capacity to recognize the truth of *G*, which he acknowledges is dependent on our recognition of the consistency of PA, we might look at the corresponding conditional claim:

(1) If PA is consistent, then *G* is not derivable in PA.

Our assertion that (1) is true is not itself crucially dependent on any assumption, such as the consistency of PA, that we are tacitly keeping

---

[1] "Can a Turing Machine Know that the Gödel Sentence Is True?" this JOURNAL, XCVI, 10 (October 1999): 525-32.

0022-362X/00/9708/456-61

from the formal system PA. But (1) will not do as an example of a truth that we can apprehend but PA cannot prove. For in PA, one can derive:

(1′) $\mathrm{Con(PA)} \rightarrow \sim\mathrm{Theorem}([G])$

(For any sentence S, $[S]$ is the formal numeral corresponding to its Gödel number. Theorem(y) is the formula $(\exists x)\mathrm{Proof}(x, y)$, where Proof($x$, y) is a formula asserting that x is the Gödel number of a proof of the sentence with Gödel number y. Thus, Theorem(y) expresses the statement that y is the Gödel number of a theorem of PA.) The conditional statement in (1′) is a formalization of part of the proof of Gödel's first incompleteness theorem. (It was the realization that this argument can be formalized which led Gödel to his second incompleteness theorem.)

McCall recognizes this, and suggests instead that we consider the following conditional claim:

(2) If PA is consistent, then $\sim G$ is not derivable in PA.

His argument has two parts: first, that (2) can be seen to be true by us and, second, that the formal counterpart of (2) cannot be derived in PA, that is, that (2′) is not a theorem of PA:

(2′) $\mathrm{Con(PA)} \rightarrow \sim\mathrm{Theorem}([\sim G])$

Therefore, he concludes, (2) finally gives us an example of a statement which we can recognize to be correct but which PA cannot prove.[2]

His argument is flawed, however. His first thesis, that (2) is correct, will come as a surprise to all those (including Gödel) who believe that establishing the nonprovability of $\sim G$ requires the stronger condition of ω-consistency.[3] McCall does offer an informal argument for (2), but he fails to realize that the argument makes tacit use of the assumption of ω-consistency. McCall argues as follows:

---

[2] McCall also finds it significant that we can recognize, again without assuming the consistency of PA, that either G is true and unprovable in PA, or G is false and provable in PA (depending on whether PA is, respectively, consistent or inconsistent). He seems to believe that this is not something a machine could establish: "Human beings, on the other hand, are acquainted with both proof and truth, and also know of cases where the two diverge" (527). Yet it is difficult to see how one's recognition of these facts could support the claim that one's mind cannot be modeled by a formal system. After all, if McCall's mind were a machine, he could have arrived at precisely the same conclusions. For in PA, one can derive $(G \,\&\, \sim\mathrm{Theorem}([G])) \vee (\sim G \,\&\, \mathrm{Theorem}([G]))$: this disjunction is equivalent to $G \leftrightarrow \sim\mathrm{Theorem}([G])$, whose derivability in PA is guaranteed by the manner in which G is constructed.

[3] To say that a theory T is ω-consistent means that there is no formula $P(x)$ such that $\vdash_T (\exists x)P(x)$—that is, $(\exists x)P(x)$ is derivable in T—but for every natural number n, $\vdash_T \sim P(\bar{n})$, where $\bar{n}$ is the numeral for n.

> Suppose, for reductio, that PA is consistent and that $\sim G$ is a theorem. Then $[\sim G]$[4] $\in$ *Th* [the collection of Gödel numbers of theorems of PA], and from the consistency of PA we infer that $[G] \notin$ *Th*. At the same time, since $\sim G$ is equivalent to $(\exists x)(\underline{\quad}x\underline{\quad}[G]\underline{\quad})$,[5] from $\vdash \sim G$ we derive $\vdash (\exists x)(\underline{\quad}x\underline{\quad}[G]\underline{\quad})$. Because the open formula $(\exists x)\underline{\quad}x\underline{\quad}y\underline{\quad}$ weakly represents *Th*, it follows that $[G] \in$ *Th*, which completes the reductio (529).

To claim that a formula $P(y)$ weakly represents *Th*, the set of Gödel numbers of theorems of PA, is to say that:

> for any n, $n \in Th$ if and only if $\vdash_{PA} P(\bar{n})$

where $\bar{n}$ is the formal numeral for the natural number n. McCall's argument relies on the correct claim that the formula $(\exists x)(\underline{\quad}x\underline{\quad}y\underline{\quad})$, that is, $(\exists x)\mathrm{Proof}(x, y)$, weakly represents the set *Th*. What he fails to realize is that *the truth of this claim depends on the assumption of ω-consistency*. In particular, the claim that $\vdash_{PA}(\exists x)\mathrm{Proof}(x, [G])$ entails that the Gödel number of G is in *Th*, which is precisely the claim needed in McCall's informal argument, depends on the ω-consistency of PA. Should PA be ω-inconsistent, it might be that $(\exists x)\mathrm{Proof}(x, [G])$ is derivable in PA even though G itself is not a theorem of PA because each n fails to be the Gödel number of a proof of G.

It is, in fact, quite clear that from the consistency of a system alone one cannot infer that the negation of its Gödel sentence is not provable. For instance, consider the formal theory $T = \mathrm{PA} + \sim\mathrm{Con}(\mathrm{PA})$, whose axioms are those of PA, together with the additional axiom $\sim\mathrm{Con}(\mathrm{PA})$. If we assume that PA is consistent, then by the second incompleteness theorem, T is as well, and clearly in T one can derive $\sim\mathrm{Con}(\mathrm{PA})$. Since an inconsistency in PA is also an inconsistency in T, $\sim\mathrm{Con}(T)$ is also derivable in T. But $\sim\mathrm{Con}(T)$ is provably equivalent to $\sim G_T$, where G_T is the Gödel sentence for T, so $\sim G_T$ is also derivable in T.

Once again, the significance of the need to assume ω-consistency may become clearer if we consider other formal systems whose ω-consistency is less evident than that of PA. McCall himself suggests (531) that we consider the formal system ZF, a formalization of Zermelo-Frankel set theory, and proposes the conditional 'If ZF is consistent then $\sim G_{\mathrm{ZF}}$ is not provable in ZF' as an example of a sentence that a human can recognize as true, but whose formal counterpart is not provable in ZF. But the claim that a human can recognize the truth of this conditional is implausible, for recognizing its truth would require knowing that ZF is ω-consistent.

---

[4] Here, McCall is using $[\sim G]$ to denote the Gödel number of the sentence $\sim G$, rather than the numeral for that Gödel number.

[5] Here, however, McCall uses $[G]$ to denote the formal numeral corresponding to the Gödel number of G. His $\underline{\quad}x\underline{\quad}[G]\underline{\quad}$ is the formula we are calling $\mathrm{Proof}(x, [G])$, and thus his sentence $(\exists x)(\underline{\quad}x\underline{\quad}[G]\underline{\quad})$ would be, in our notation, $\mathrm{Theorem}([G])$.

Because of this error, McCall's claim that PA cannot derive (2′) is quite irrelevant. Should the claim be correct, it would not establish any difference between humans and machines, for humans are just as incapable of establishing (2).

As it happens, the claim is correct, although McCall fails to see why. He offers a kind of plausibility argument to the conclusion that the derivation in PA of (2′) is "highly unlikely" (529). But, in fact, there is a straightforward proof that, if PA is ω-consistent, then (2′) is not derivable in PA. Assume that $\mathrm{Con}(\mathrm{PA}) \rightarrow \sim\mathrm{Theorem}([\sim G])$ is derivable in PA. Since Con(PA) is provably equivalent to G, it follows that $\mathrm{Theorem}([\sim G]) \rightarrow \sim G$ is also derivable in PA. But by Löb's Theorem,[6] it follows that $\sim G$ is likewise derivable in PA, and so PA is not ω-consistent. Hence, if PA is ω-consistent, then (2′) is not derivable in it.

We have shown that what humans are actually capable of establishing is:

(3) If PA is ω-consistent, then $\sim G$ is not derivable in PA.

There is still the possibility that McCall's argument can be rescued by showing that *this* conditional is one that humans can recognize to be true but PA cannot derive. Specifically, one might wonder whether:

(3′) $\omega\text{-}\mathrm{Con}(\mathrm{PA}) \rightarrow \sim\mathrm{Theorem}([\sim G])$

is derivable in PA. Another way of putting the question is this: If PA is given the same information that McCall tacitly makes use of in establishing (2), can it derive (2′)?[7]

---

[6] Löb's Theorem says that for any sentence S, if $\vdash_{\mathrm{PA}} \mathrm{Theorem}([S]) \rightarrow S$, then $\vdash_{\mathrm{PA}} S$.

[7] One might also wonder whether McCall's argument can be fixed by using Rosser's sentence, rather than Gödel's sentence G, in sentence (2) and its formal counterpart (2′). For sentence (2), so modified, can be established without relying on a tacit assumption that PA is ω-consistent. It can be shown, however, that the modified version of (2′) is derivable in PA. Thus, this strategy would not produce McCall's sought-after distinction between mind and machine.

The answer is "Yes." Here is one proof: we begin by observing that, if something fails to be the Gödel number of a proof, then, since this can be determined by a finite check, it is provable that it fails to be. In fact, this observation can itself be proven in PA, since for any sentence S:

(4.1) $\vdash_{\mathrm{PA}} (\forall n)(\sim\mathrm{Proof}(n, [S]) \rightarrow \mathrm{Theorem}(\mathrm{sub}(n, [\sim\mathrm{Proof}(x, [S])])))$

where $\mathrm{sub}(n, [P])$ denotes the Gödel number of the result of substituting the formal numeral for n for all free variables in the formula P.

According to the definition of 'Theorem(y)':

(4.2) $\vdash_{\mathrm{PA}} \mathrm{Theorem}([\mathrm{Theorem}([S])]) \rightarrow \mathrm{Theorem}([(\exists x)\mathrm{Proof}(x, [S])])$

We also know that:

(4.3) $\vdash_{\mathrm{PA}} \sim\mathrm{Theorem}([S]) \rightarrow (\forall n)\sim\mathrm{Proof}(n, [S])$

Applying (4.1), we get:

(4.4) $\vdash_{\mathrm{PA}} \sim\mathrm{Theorem}([S]) \rightarrow (\forall n)\mathrm{Theorem}(\mathrm{sub}(n, [\sim\mathrm{Proof}(x, [S])]))$

Combining (4.2) and (4.4), we arrive at:

(4.5) $\vdash_{\mathrm{PA}} (\mathrm{Theorem}([\mathrm{Theorem}([S])]) \,\&\, \sim\mathrm{Theorem}([S])) \rightarrow$

$(\mathrm{Theorem}\ ([(\exists x)\mathrm{Proof}(x, [S])]) \,\&$

$(\forall n)\mathrm{Theorem}\ (\mathrm{sub}(n, [\sim\mathrm{Proof}(x, [S])])))$

But the right-hand side of (4.5) is exactly what ω-consistency rules out. Hence:

(4.6) $\vdash_{\mathrm{PA}} (\mathrm{Theorem}([\mathrm{Theorem}([S])]) \,\&\, \sim\mathrm{Theorem}([S])) \rightarrow \sim\omega\text{-}\mathrm{Con}(\mathrm{PA})$

In other words:

(4.7) $\vdash_{\mathrm{PA}} \omega\text{-}\mathrm{Con}(\mathrm{PA}) \rightarrow (\mathrm{Theorem}([\mathrm{Theorem}([S])]) \rightarrow \mathrm{Theorem}([S]))$

Note that (4.7) formalizes half of the claim that the ω-consistency of PA implies that the formula Theorem(y) weakly represents *Th*. In particular, it is the formal counterpart of the claim that, if PA is ω-consistent, then, if $\vdash_{\mathrm{PA}}\mathrm{Theorem}([S])$, then the Gödel number of S is an element of *Th*.

We now apply (4.7) to the formula G to get:

(4.8) $\vdash_{\mathrm{PA}} \omega\text{-}\mathrm{Con}(\mathrm{PA}) \rightarrow (\mathrm{Theorem}([\mathrm{Theorem}([G])]) \rightarrow \mathrm{Theorem}([G]))$

Since G is provably equivalent to $\sim\mathrm{Theorem}([G])$, this implies that:

(4.9) $\vdash_{\mathrm{PA}} \omega\text{-}\mathrm{Con}(\mathrm{PA}) \rightarrow (\mathrm{Theorem}([\sim G]) \rightarrow \mathrm{Theorem}([G]))$

Of course, by the definition of consistency we have:

(4.10) $\vdash_{\mathrm{PA}} \mathrm{Con(PA)} \rightarrow (\mathrm{Theorem}([\sim G]) \rightarrow \sim\mathrm{Theorem}([G]))$

Combining (4.9) and (4.10), and using the fact that $\vdash_{\mathrm{PA}} \omega\text{-}\mathrm{Con(PA)} \rightarrow \mathrm{Con(PA)}$, we reach the desired conclusion:

(4.11) $\vdash_{\mathrm{PA}} \omega\text{-}\mathrm{Con(PA)} \rightarrow \sim\mathrm{Theorem}([\sim G])$

In short, if we level the playing field between mind and machine, they remain in a dead heat. At least, McCall has not succeeded in showing otherwise.

ALEXANDER GEORGE

Amherst College

DANIEL J. VELLEMAN

Amherst College

COMMENTS AND CRITICISM

WHAT GÖDEL'S INCOMPLETENESS RESULT DOES AND DOES NOT SHOW

In a recent paper, Storrs McCall[1] adds another link to a chain of attempts to enlist Kurt Gödel's incompleteness result as an argument for the thesis that human reasoning cannot be construed as being carried out by a computer. McCall's paper is undermined by a technical oversight. My concern however is not with the technical point. The argument from Gödel's result to the no-computer thesis can be made without following McCall's route; it is then straighter and more forceful. Yet the argument fails in an interesting and revealing way. And it leaves a remainder: if some computer does in fact simulate all our mathematical reasoning, then, in principle, we cannot fully grasp how it works. Gödel's result also points out a certain essential limitation of self-reflection. The resulting picture parallels, not accidentally, Donald Davidson's view of psychology as a science that in principle must remain "imprecise," not fully spelled out. What is intended here by 'fully grasp', and how all this is related to self-reflection, will become clear at the end.

I should add that the full implications and the significance of Gödel's result are often misunderstood in other important respects. The result is rightfully conceived as revealing an essential limitation of formal deductive systems: in any deductive system that satisfies the non-negotiable requirement of effectiveness (the existence of an effective procedure for deciding whether any purported proof is a proof), and some minimal adequacy conditions (the capacity to represent certain elementary arithmetic, or combinatorial notions), there are sentences that are neither provable nor refutable in the system. What this formulation hides, and what the proofs current in most textbooks miss, is the constructive nature of the original proof, which does not appeal to the truth concept. It is by virtue of this feature that the result has undermined Hilbert's finitistic program in the foundation of mathematics. Although Gödel's theorem is an independence result, it is not a radical one, like the independence of

---

[1] "Can a Turing Machine Know that the Gödel Sentence Is True?" this JOURNAL, XCVI, 10 (October 1999): 525-32. The two previous most noted attempts have been John Lucas, "Minds, Machines and Gödel," *Philosophy*, XXXVI (1961): 112-27, and Roger Penrose in *The Emperor's New Mind* (New York: Oxford, 1989) and *Shadows of the Mind* (New York: Oxford, 1994).

0022-362X/00/9708/462-70

the continuum hypothesis; rather the opposite: we know, by good reasons as any, the truth value of the independent sentence. And it is this "weak" character which endows Gödel's result with its foundational significance and with its force against Hilbert's program. Finally, while the theorem has profound effects on the foundations of mathematics and on our view of formal systems, it has had very little effect on mathematical practice. These and other related points are beyond the scope of the present reflection; I hope to address them elsewhere.

Let me first refine, for clarification, what I called above the *no-computer thesis.* The claim is that no computer program that produces, or searches for, proofs can achieve what is achievable by mathematicians who engage in proving theorems by mathematical reasoning. This thesis is of great interest in itself, and I shall not be concerned here with the broader implications it might have in the philosophy of mind. The argument for the thesis consists in the following two moves. (I) A theorem prover (that is, a computer program that proves theorems) can be seen as finding theorems in some formal deductive system. (II) For any formal deductive system there is a sentence (for example, the so-called Gödel sentence that says of itself that it is unprovable in the system) which cannot be proved in the system but whose truth can be inferred by mathematical reasoning. (I) is not problematic. A theorem prover generates a recursively enumerable set of sentences (the general reader may ignore this technical detail without missing much), and the closure of this set under logical implications can be seen as the set of theorems of some formal deductive system (we can actually assume that the theorem prover generates the closure of this set). My concern here will be with (II). At this point a minimal sketch of Gödel's proof may not be redundant, given that some elementary misunderstandings can still be found in the philosophical literature.[2]

Usually, Gödel's result is stated for number-theoretic systems, and its proof uses an encoding of sentences, sentence parts, and proofs by natural numbers: the so-called Gödel numbers. This makes it possible to express within a formal theory of numbers the basic syntactic, and proof-theoretic, notions of the system itself. But the result can be stated and proved with respect to any system that can handle its own syntax—for example, some theory of finite strings. Let T (referred to also as 'theory') be a system based on a first-order language, whose intended interpretation is over the universe of natural numbers. (The universe may also include other objects, provided that the subclass of natural numbers can be defined by a suitable formula in the language.) The language has the usual vocabulary: names for addition and multiplication, a name, $\underline{0}$, for the number 0, and a name, $s(\ )$, for the successor function. Standard names for other numbers are obtained, in the usual way, by iterating $s(\ )$; for example, the name, $\underline{4}$, for the number 4 is $s(s(s(s(\underline{0}))))$. The axioms should be sufficient for deriving certain basic properties, which enable one to represent various syntactic notions inside the system, where formulas and proofs are "identified with" their Gödel numbers. This, it turns out, is a rather weak requirement, which is satisfied by systems considerably weaker than Peano's arithmetic.[3] Under these conditions, there is a formula: $Proof_{\mathrm{T}}(y,x)$, which says that y is a proof, in T, of x, such that the following holds (where 'T ⊢ ...' means that, in T, ... is provable, $\underline{n}$ is the standard name of n, and proofs and sentences are identified with their Gödel numbers).

---

[2] Daniel C. Dennett, for example, in chapter 13 of *Brainstorms* (Cambridge: MIT, 1986) (I thank Ihsan Dogramachi for calling my attention to this chapter), speaks of Gödel sentences of Turing machines, including (on page 265) the Gödel sentence of the universal Turing machine. This makes no sense. A Gödel sentence is defined with respect to a deductive system, or with respect to a theorem prover, which a Turing machine is in general not. "The Gödel sentence of the universal Turing machine" is either meaningless, or—if by stretch of imagination the "universal Turing machine" is supposed to cover all deductive systems—a false, or even contradictory sentence.

(1) *If* n is a proof, in T, of m, then $\mathrm{T} \vdash Proof_{\mathrm{T}}(\underline{n},\underline{m})$

(2) *If* n is not proof, in T, of m, then $\mathrm{T} \vdash \neg Proof_{\mathrm{T}}(\underline{n},\underline{m})$

At the heart of Gödel's proof is the fixed-point theorem, also known as the diagonal theorem, which says that, for any given formula, $\phi(x)$, with one free variable, there is a sentence α such that the following biconditional is provable in T:

$$\alpha \leftrightarrow \phi(`\alpha')$$

where 'α' is the standard name of α's Gödel number. (Moreover, the proof of the theorem yields an actual construction of α, as well as of the proof of the biconditional). We can think of α as a sentence that, referring to itself via its Gödel number, says of itself that it has the property $\phi(\ )$. Letting $\phi(x)$ be the formula $\neg\exists y Proof_{\mathrm{T}}(y,x)$, we get a sentence γ that says of itself that it is not provable (in T):

---

[3] For example, it is sufficient that the following be provable: all true inequalities $\underline{m} \neq \underline{n}$, all true equalities $\underline{m} + \underline{n} = \underline{k}$, and similarly for multiplication; also, for each n, the formulas: $x < \underline{n} \leftrightarrow x = \underline{0} \vee x = \underline{1} \vee ... \vee x = \underline{n\text{-}1}$ and $x < \underline{n} \vee x = \underline{n} \vee \underline{n} < x$ (where the ordering, <, is expressed either through some formula, or is named by a relation symbol of the vocabulary).

(3) $T \vdash \gamma \leftrightarrow \ \exists y Proof_{\mathrm{T}}(y, `\gamma')$

Now, if n is any proof (in T) of γ, then, combining it with the proof of the biconditional, we get a proof of $\neg\exists y Proof_{\mathrm{T}}(y,x)$; on the other hand, (1) gives us a proof of $Proof_{\mathrm{T}}(\underline{n}, `\gamma')$. Since these two sentences contradict each other (in first-order logic), we get a proof of a contradiction. Therefore, if T is consistent, $\gamma$ is not provable in it. But the unprovability of $\gamma$ is formally stated as $\neg\exists y Proof_{\mathrm{T}}(y, `\gamma')$; hence, we have just inferred $\neg\exists y Proof_{\mathrm{T}}(y, `\gamma')$ from the assumption that T is consistent. Combining this with the biconditional in (3), we infer γ. Now the claim that T is consistent can be formalized in T, say, as the sentence that says that there is no proof of $\underline{0} \neq \underline{0}$: $\neg\exists y Proof_{\mathrm{T}}(y, `\underline{0} \neq \underline{0}')$. Call this sentence Con(T). Then the argument we have just given amounts to a derivation (outside T) of:

(4) $Con(T) \rightarrow \neg\exists y Proof_{\mathrm{T}}(y, `\gamma')$

If T is minimally adequate, this derivation can be formalized in it, making (4) a theorem of T. ('Minimally adequate' is still rather weak, considerably less than Peano's arithmetic.) Assuming from now on that T is minimally adequate, it follows that, if T is consistent, Con(T) is not provable in it; otherwise, we would get from (4) a proof of $\neg\exists y Proof_{\mathrm{T}}(y, `\gamma')$, hence a proof of γ. This is the content of Gödel's second incompleteness theorem.

The reason we accept the Gödel sentence, γ, as true is that it is implied by the consistency of T. Rather than γ, let us therefore focus on Con(T) as the unprovable sentence. (Indeed, γ is provably equivalent to Con(T): we have just indicated how to get γ from Con(T) and, in the other direction, Con(T) follows from the unprovability of any sentence, in particular from $\neg\exists y Proof_{\mathrm{T}}(y, `\gamma')$.) The basis of all philosophical take-offs from Gödel's result is thus the following simple fact: for any minimally adequate formal deductive system, T, if T is consistent then its being consistent is expressible in T but unprovable in it.

Before proceeding, a short remark on McCall's oversight. Gödel's theorem has also a second part, proved by similar arguments, which says that, if T is ω-consistent, then $\neg\gamma$ is not provable in it; ω-consistency is the following condition that is stronger than consistency: whenever $\exists x \alpha(x)$ is provable (where 'x' ranges over the natural numbers), it is not the case that $\neg\alpha(\underline{n})$ is provable for all n. McCall observes that it is unknown whether ω-consistency can be replaced in the second part by consistency. In fact, it is not difficult to show that it *cannot*, unless T is not ω-consistent. But all this is neither here nor there, for the use of ω-consistency does not make for a new or improved argument. We can express in the language of T the statement that T is ω-consistent and get an analogous conditional to (4): $\omega\text{-Con}(\mathrm{T}) \rightarrow \neg\exists y \mathit{Proof}_{\mathrm{T}}(y, \ulcorner\neg\gamma\urcorner)$. This, again, is provable in T, under mild adequacy assumptions. The situation is similar to that of the first part. If T is ω-consistent, this fact cannot be proved in it. Since consistency is a weaker and more natural property than ω-consistency, the argument for the no-computer thesis is better if it is based on the unprovability of Con(T). Let me now turn to this argument.

Any deductive system, T, that formalizes mathematical reasoning must leave something outside: its own consistency, expressed as Con(T), cannot be derived in it. As we remarked above, in principle, a computer that proves theorems generates proofs in some formal system. (This is true of the standard notion of computers. It may fail of computers that incorporate physical elements that, by virtue of physical theory, compute nonrecursive functions. I shall not pursue this possibility here.) If the computer can "know" only what it can prove, then it cannot know that it is consistent (that is, never produces a contradiction), something that we qua mathematicians seem to know. We may know it, for example, on the basis of a soundness argument, by appealing to the concept of truth: all the axioms of T are true, in the intended interpretation of the language in question, and all the inference rules preserve truth; therefore, whatever we prove from the axioms by applying inference rules, must be true, a fortiori, noncontradictory.

There is now a philosophical temptation to argue that our access to the concept of truth gives us an edge over any computer. The thought appears attractive, given the traditional distinction between syntax and semantics and the view that sees the computer as carrying out mere syntactic manipulations. But this line fails. In the same sense that computers can prove, they can handle semantic notions. The argument from soundness can be formalized in a richer system, in which we can define a truth predicate for the language of T; the clauses of the truth definition are deducible as theorems in the second system, and so is the general statement that every sentence provable in T is true, hence noncontradictory. A computer that proves theorems in the richer system "knows" that T is consistent.

The appeal to the truth concept is therefore misplaced. Still, the argument seems to carry through. As a rule, mathematicians take the consistency of the framework within which they work for granted. (This is different from physics where *faute de mieux* contradictions are tolerated as temporary anomalies, to be resolved later.) If T is obtained by formalizing a framework of mathematical reasoning, those who accept the framework accept, at least implicitly, that T is consistent; that is, they accept Con(T). This acceptance is not merely some state ascribed to the mathematician for philosophical convenience. Had it been so, one might have stipulated "by definition" that the theorem prover that works in T "knows" that T is consistent. In principle, the acceptance can be manifested by using Con(T) in mathematical practice; a mathematician will reject as false any hypothesis he is trying to establish, if he finds that the hypothesis implies the inconsistency of the system within which he is working. Which is what a theorem prover is incapable of. Observe, moreover, that the mathematician's ability derives from a certain capacity for self-reflection. Reflecting on one's mathematical reasoning and realizing that it can be formalized as T, one takes Con(T) as something that can be legitimately used in mathematical reasoning.

We can extend T by adding Con(T) as an additional axiom. But then the consistency of *that* system is beyond *its* reach, and so on. We thus have a systematic way of extending any acceptable formal system to a stronger acceptable system. That at any single stage we can view ourselves as generating proofs in some formal system does not invalidate the no-computer argument, for the fact remains that no single system, hence no single computer, can faithfully capture what we can do. Or does it?

A closer look reveals the hole in this line of thought. Our alleged edge over the computer that generates proofs in T consists of our knowing that T is consistent. But how do we know this? If I believe that my mathematical reasoning is immune to contradiction, and *if I recognize that the formal system* T *represents faithfully my reasoning,* then indeed I should accept Con(T). But T may represent my reasoning without me recognizing this fact. Call a formal system (theory) *recognizably valid* if it can be recognized as a theory whose theorems should be accepted, because it represents faithfully our mathematical reasoning (or because it derives from some such representations). Then, a recognizably valid T cannot capture our full mathematical reasoning; because we shall reason that T is consistent, and *this* reasoning is something T cannot deliver. Yet, for all we know, there may be a system, call it T*, which is not recognizably valid and which in fact represents, in a very complicated way, all our mathematical reasoning. A computer that generates T*-proofs simulates all our doings in mathematics, in all the systems we can create. Suppose we are told that a certain complicated theorem prover, consisting of millions of instructions, produces exactly all that human mathematical thinking is capable of. Whoever believes it will probably conclude that the algorithm is consistent (cannot produce a contradiction). But what grounds can we have for such a belief? Say we believe it on the authority of some superior being (create, dear reader, your own favorite story). Or say we believe it on the inductive force of empirical evidence: somehow we got this program (again, create your own scenario), which has proved all major mathematical theorems known to us, and whose output, as far as we checked it, contained no errors. But such a belief does not derive from *mathematical* reasoning; hence it is not a counterexample to the computer's alleged ability to simulate human mathematical reasoning.

One should note here that the ability to form beliefs on someone's word, or on empirical evidence, does not by itself constitute a principled difference between humans and computers. For a computer can be provided with channels through which it is continuously fed environmental input, which can affect its deductive processes. It is not clear to what extent the no-computer argument can be extended to this case (by using a suitable version of the diagonal theorem). In any case, this goes beyond the topic of Gödel's incompleteness result. Our present discussion does however cover the case of a mathematical community, whose members exchange proofs and ideas. For we can model such a community as a network of intercommunicating computational modules, a system that constitutes one giant computer program. The particulars of the program do not matter, as long as the total output is recursively enumerable, which it will be if each module behaves according to some algorithm. The same goes for a single human, who can be construed as performing several tasks in parallel; there will be several subroutines that form one overall program.[4] None of this affects the arguments.

Now in order to believe the consistency of some program on *mathematical* grounds, we have to understand how the program works, just as we understand that a certain algorithm performs addition, or just as we understand that Peano's axioms are true in the model of natural numbers and that the inference rules preserve truth. Yet, when it comes to T* such understanding is beyond our capacity; the system, or the computer, is too complex for us to grasp as an object of mathematical thought. This is not merely a question of length or of our memory getting worn down, but a limit that exist in principle, since it is implied by Gödel's theorem. (It is similar to upper and lower bounds that are established in complexity theory.)

---

[4] Dennett, *op. cit.*, tries to defuse the implications of the no-computer thesis by arguing that human behavior may represent several Turing machines, running in parallel, such that the Gödel sentence that is unprovable by one is provable by another. The argument, it can now be seen, misfires. If the total output is recursively enumerable, we have in principle one program. And if it is not, then it remains for the functionalist to explain how mathematical reasoning, performed according to some functionalist model, produces a nonrecursively enumerable set. Dennett also argues that the same physical device can implement, according to different interpretations, many different programs. This is trivially true, but of no help to the functionalist, who is after all committed to there being a particular computational model that explains mental activity.

Physical limitations on our performance exist, of course, in general. Although we have a good grasp of the formal system of Peano's arithmetic and recognize it as valid, certain deductions in it are beyond us, because of memory wear, or shortage of time. Still we idealize and say that, were it not for these limitations, humans could prove each theorem of Peano's arithmetic. When it comes to T*, we can say that the theorems of T* coincide with what is humanly provable in principle, that is, disregarding physical limitations on human performance. But even under these idealized conditions, we could not, as a matter of principle, recognize T* as valid. This would be a structural inherent feature, rather than a matter of wear and tear.

And if such is the case, then we (qua mathematicians) are machines that are unable to recognize the fact that they are machines. As the saying goes: if our brains could figure out how they work, they would have been much smarter than they are. Gödel's incompleteness result provides in this case solid grounds for our inability, for it shows it to be a mathematical necessity. The upshot is hauntingly reminiscent of Spinoza's conception, on which humans are predetermined creatures, who derive their sense of freedom from their incapacity to grasp their own nature. A human, namely, Spinoza himself, may recognize this general truth; but a human cannot know *how* this predetermination works, that is, the full theory. Just so, we can entertain the possibility that all our mathematical reasoning is subsumed under some computer program; but we can never know how this program works. For if we knew we could diagonalize and get a contradiction.

The discussion above brings to the fore the role of mathematical self-reflection: A mathematician realizes by self-reflecting on his own reasoning that his inferences can be formalized by such-and-such deductive system. From which the mathematician can go on to infer that the system in question is consistent. The self-reflection is therefore itself part of mathematical reasoning. Gödel's result shows however that self-reflection cannot encompass the whole of our reasoning; that is, it cannot comprehend itself within its horizon. There is, indeed, prima facie plausibility to the general claim that we can reflect on only part of our rational apparatus; for the very act of self-reflection remains outside the picture it reveals. Gödel's result lends mathematical grounds to this intuitive plausibility.

Finally, note how the impossibility of full grasp of our own "reasoning program" accords, in spirit, with Davidson's claim that psychology must remain on a certain level of imprecise generalities. We may speculate how our reasoning works and we may confirm some general aspects of our speculation. But we cannot have a full detailed theory. The reason for the impossibility is the same, both in the case of mathematical reasoning and in the case of psychology, namely, the theoretician who constructs the theory is also the subject the theory is about.

HAIM GAIFMAN

Columbia University

BOOK REVIEWS

Sex and Social Justice. MARTHA NUSSBAUM. New York: Oxford University Press, 1999. ix + 476 p. Cloth $35.00.

The final essay in this impressive volume is a discussion of the quotidian yet complex interpretive enterprise of understanding other people. It presents, in domestic microcosm, many of the themes treated at a more general level in the preceding essays. The focus is on the couple at the center of Virginia Woolf's *To The Lighthouse*, who communicate without giving voice, and concede or hold their ground over their differences without making anything explicit. They sustain a subtle, if fallible, working understanding of each other's thoughts and experiences by drawing silently on their intimate history together. Although most of the rather noisier contexts in which human beings may try to understand each other do not involve any such history of intimacy, Nussbaum's concluding essay appropriately directs one's attention backward through the volume so that one momentarily rereads it whole as an elaboration on our human capacity for sympathetic, sometimes critical, understanding—not only of other individuals but also of other cultures. One must avoid romanticism, of course. Interpersonal understanding between individuals who share a way of life is one thing, and crosscultural understanding another. The question of crosscultural understanding tends to present itself in Western thought as either a hope or an anxiety as to whether we may succeed in understanding others well enough to avoid doing them any harm. (Thinking from a position of power means that the question whether others may be able to understand *us* well enough so that we might avoid being harmed tends not to be raised.) Whether in the form of anxiety or hope, the philosophical question demands attention; but it must not be allowed to paralyze first-order ethical thought. Thus it is right that Nussbaum should address the issue principally at the first-order level. Not, then, by engaging in extended a priori argument about the very idea of crosscultural understanding, but rather through paying close attention to specific cases (the most relevant essay in this connection perhaps being "Judging Other Cultures: The Case of Genital Mutilation"). The task is to look and see whether or not the response that a certain practice is unjust constitutes a sensitively informed judgment, or merely an ethical-imperialist reaction. If some particular practice deprives girls and women of realizing a fundamental human

0022-362X/00/9708/471–75

capability (the capability for bodily integrity, for example), then one will have found that something humanly universal is offended and the practice is unjust. The idea of a common humanity is at the heart of Nussbaum's position, and is specified as a set of basic human functional capabilities. The capabilities approach (which, as she says, has also been developed by Amartya Sen) is really the starting point of the book, and the explicit subject matter of the first essays, in which she presents and defends her universalist feminist position.

On Nussbaum's view, there are a number of basic functional capabilities possessed universally by human beings, which justice requires must not be thwarted. If the universalism is to be sound, these capabilities must be described so that they are constant through historical and cultural diversity. But then the list of "Central Human Functional Capabilities" as it is presented here includes two or three things that do not belong. It includes capabilities described in a way that makes them dependent on culturally and historically contingent institutions. The third entry on the list includes "marital rape," which presupposes an institution of marriage; the tenth includes "being able to hold property" and "the right to seek employment on an equal basis with others," which presuppose institutions of private property and competitive employment which belong specifically to capitalist economic systems. These are glitches in an otherwise very plausible and ethically powerful list. But I think they are worth noting, not because they must cast doubt on the overall project—they could surely be reexpressed in more neutral terms, or, if not, the list would in any case remain substantial enough without them—but rather because confidence in the universalist project must be earned precisely by showing that historical and cultural specificity can be kept from creeping into the description of the purportedly universal capabilities.

Nussbaum's feminism "has five salient features: It is *internationalist, humanist, liberal, concerned with the social shaping of preference and desire,* and, finally, *concerned with sympathetic understanding*" (6). It is interesting to reflect on how these features fit together. Many of the essays here contain empirical data and individual case studies (Nussbaum worked for some years for a development agency connected with the United Nations). The empirically informed and politically engaged character of her internationalism affords the theoretical features of her approach their full weight in a deeply convincing ethical position. The engaged stance is a particular strength in staving off the automated skeptical responses of those in the grip of an a priori skepticism about universalist or (as Nussbaum uses the term) humanist projects. One of the conclusions to be drawn from this book, I think, is that any responsible skepticism to the effect that such projects are inevitably culturally imperialist will not be a skepticism a priori. Philosophical skepticism about the very idea of universal human capabilities, or of universal rights to safeguard them, is intellectually facile and ethically obnoxious when clung to in the face of social realities in which women are asymmetrically deprived of the basic education they need for minimal material independence (30), or in which they are prohibited from working outside the home even when staying there means virtual starvation (29). Rather, if there is a responsible skepticism to be reckoned with here, it will be more a historical than a philosophical one, consisting principally of a worldly pessimism at anyone's (and perhaps especially the West's) chances of managing to shape a universalism that avoids doing violence to cultural difference. But a skepticism of pessimism, combined with a practical commitment to justice, is not something to make one give up on a universalism of basic human capabilities and their defense. On the contrary, it is something to make one pursue it with careful and self-conscious attention to particulars—hence the importance of the concern with sympathetic understanding.

The connection between Nussbaum's liberalism and the concern with the social shaping of preference and desire is particularly interesting, not least because the connection is surely not quite as presented. An awareness of the power of ideology (if I may put it that way) in shaping preferences serves as an important corrective to liberalism's asocializing individualist drift. As Nussbaum nicely puts it:

> Any living culture contains plurality and argument; it contains relatively powerful voices, relatively silent voices, and voices that cannot speak at all in the public space. Often some of these voices would speak differently, too, if they had more information or were less frightened—so part of a culture, too, is what its members *would* say if they were freer or more fully informed (8).

Quite. As I see it, then, Nussbaum's liberalism takes from feminism what feminism took from socialism: most particularly, the insight that what oppressed people want is not always what is in their interests, that is, not always what they would want if they had access to the bigger picture. One of the essays is entirely devoted to the feminist critique of liberalism, and Nussbaum's espousal of liberalism is explicitly informed by that critique—specifically, by the idea that desires and preferences, and indeed emotions quite generally, are socially formed. But it is odd that the only precedents traced for this idea are precursors in the liberal tradition. It is well worth having it drawn to one's attention that Jean-Jacques Rousseau, Adam Smith, and J. S. Mill (78) all expressed views about the social shaping of emotions, and that so did the Greek and Roman Stoics (Introduction, 12). But the liberalism that Nussbaum espouses surely owes a direct debt to socialist thought—most notably the insights expressed there in terms of ideology and false consciousness—and this debt is obscured by the emphasis on liberal and ancient precursors. Nussbaum's feminism, then, surely has a more mixed intellectual inheritance than she represents. But the result is no less exciting: a liberal feminist position which (like some other liberal feminist work—for instance, that of Susan Okin) is properly inflected by an awareness that inequalities of power can create an uneven pitch for the actual exercise of individual freedoms.

This is a wonderfully diverse volume of essays, so much so that one is forced to discuss it in the broad; and yet one cannot hope to do justice to its breadth, or its depth, by attending to the general ethical trajectory alone. There is some slight repetition, chiefly in discussions of Catharine MacKinnon and Andrea Dworkin. This is perhaps inevitable in essays written over a span of seven years, and would not be worth mentioning except that it is also in part the result of what I am personally inclined to consider an unduly exclusive attention to MacKinnon's and Dworkin's work (the exclusivity can create the false impression that the feminist insight into power's conditioning of sexual desire is owed uniquely to them; see 78). Be that as it may, this volume is a feast, with essays treating matters arising directly from development work, from the reading of fiction, from first-order political commitments, from competing philosophical conceptions of justice, and also from more personal autobiographical concerns. Thus, there are discussions of women's literacy, genital mutilation, the moral authority of religion, the place of mercy within justice, lesbian and gay rights, prostitution, and sexual objectification; and there is also a personal discussion of Kenneth Dover's life, to whom the volume is dedicated. All make enormously rewarding reading, and the philosophy is lit up by Nussbaum's sensitive and perceptive drawing on real-life stories. The two parts of the book—"Justice" and "Sex"—give a broad division of themes, but the essays are bound together as a whole by a family of thematic relations: sex, gender, sexuality, and justice. What emerges from these luminous writings is a sense of an utterly authentic ethical outlook whose commitment to the dignity of human life, in all its diverse cultural forms, may well be enough to put paid to the skepticism of pessimism with which some readers, not altogether without reason, might approach it.

MIRANDA FRICKER

Heythrop College/University of London

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Introduction to Sociology.* CHRISTOPH GÖDDE, ed., EDMUND JEPHCOTT, trans. Stanford: University Press, 2000. x, 198 p. Cloth $39.50; *Problems of Moral Philosophy.* THOMAS SCHRÖDER, ed., RODNEY LIVINGSTONE, trans. Stanford: University Press, 2000. viii, 224 p. Cloth $39.50.

AGAMBEN, GIORGIO: *Remnants of Auschwitz: The Witness and the Archive.* DANIEL HELLER-ROAZEN, trans. Cambridge: MIT, 1999. 1, 175 p. Cloth $25.00.

AQUINAS, THOMAS: *Disputed Questions on Virtue.* RALPH MCINERNY, trans. South Bend, IN: St Augustine's, 1999. xix, 140 p. Cloth $24.95.

BACON, FRANCIS: *The New Organon.* LISA JARDINE and MICHAEL SILVERTHORNE, eds., MICHAEL SILVERTHORNE, trans. New York: Cambridge, 2000. xxxv, 252 p. Cloth $59.95, paper $18.95.

BENJAMIN, WALTER: *The Arcades Project.* HOWARD EILAND and KEVIN MCLAUGHLIN, trans. Cambridge: Harvard, 1999. xiv, 1073 p. Cloth $39.95.

BOURDIEU, PIERRE: *Pascalian Meditations.* RICHARD NICE, trans. Stanford: University Press, 2000. vii, 256 p. Cloth $49.50, paper $18.95.

BULGAKOV, SERGEI: *Philosophy of Economy.* CATHERINE EVTUHOV, trans. New Haven: Yale, 2000. vii, 347 p. Cloth $35.00.

CICERO: *On the Commonwealth* and *On the Laws.* JAMES E. G. ZETZEL ed. and trans. New York: Cambridge, 1999. xlviii, 207 p. Cloth $ 54.95, paper $19.95.

DESCARTES, RENÉ: *Philosophical Essays and Correspondence,* ROGER ARIEW, ed., various trans. Indianapolis: Hackett, 2000. xxiv, 334 p. Cloth $37.95, paper $14.95.

GADAMER, HANS-GEORG: *Hermeneutics, Religion, and Ethics.* JOEL WEINSHEIMER, trans. New Haven: Yale, 1999. xiv, 172 p. Cloth $25.00.

GENETTE, GÉRARD: *The Aesthetic Relation.* G. M. GOSHGARIAN, trans. Ithaca: Cornell, 1999. xi, 243 p. Cloth $49.95, paper $22.50.

HEIDEGGER, MARTIN: *Contributions to Philosophy (From Enowning).* PARVIS EMAND and KENNETH MALY, trans. Bloomington: Indiana UP, 1999. xlvii, 369 p. Cloth $39.95.

HUMBOLDT, WILHELM VON: *On Language: On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species.* MICHAEL LOSONSKY, ed., PETER HEATH, trans. New York: Cambridge, 1999. xli, 296 p. Cloth $59.95, paper $19.95.

LAFONT, CRISTINA: *The Linguistic Turn in Hermeneutic Philosophy.* JOSÉ MEDINA, trans. Cambridge: MIT, 1999. xviii, 377 p. Cloth $45.00.

PLATO: *Statesman.* CHRISTOPHER J. ROWE, trans. Indianapolis: Hackett, 1999. xxi, 97 p. Cloth $29.95, paper $7.95; *Clotophon.* S. R. SLINGS, trans. New York: Cambridge, 1999. xv, 360 p. Cloth $69.95; *Timaeus.* DONALD J. ZEYL, trans. Indianapolis: Hackett, 2000. xcv, 94 p. Paper $12.95.

RICOEUR, PAUL: *The Just.* DAVID PELLAUER, trans. Chicago: University Press, 2000. xxiv, 161 p. Cloth $20.00.

NEW FROM INDIANA!

THE GOOD, THE BAD, AND THE BEAUTIFUL

Discourse about Values in Yoruba Culture

Barry Hallen

"Hallen renders a great service to African philosophy." —Kwasi Wiredu

Reveals everyday language as the key to understanding morals and ethics in Yoruba culture.

January 2001 256 pp
cloth $39.95 paper $17.95

THE OTHER HUSSERL

The Horizons of Transcendental Phenomenology

Donn Welton

"Welton has an incredible grasp of the whole of Husserl's work . . . insightful and detailed, yet at the same time readable and accessible."
—David Carr

An original and comprehensive reconstruction of Husserl's phenomenological method.

Studies in Continental Thought
December 2000 512 pp
cloth $49.95

AFRICAN PHILOSOPHY AS CULTURAL INQUIRY

Edited by Ivan Karp and D. A. Masolo

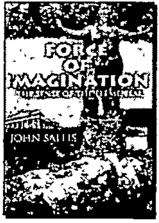
Assesses the direction and impact of African philosophy as well as its future role.

African Systems of Thought
Published in association with the International African Institute, London
December 2000 272 pp, 10 b&w photos
cloth $39.95 paper $18.95

FORCE OF IMAGINATION

The Sense of the Elemental

John Sallis



". . . a powerful, original approach to what others call 'ecology' but what Sallis shows to be a question of the status of the earth in philosophical thinking at this historical moment."
—Edward S. Casey

Studies in Continental Thought
October 2000 272 pp
cloth $39.95 paper $24.95

Now in a new edition!

REMEMBERING

A Phenomenological Study, Second Edition

Edward S. Casey

A *Choice* Outstanding Academic Book

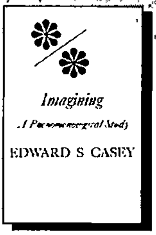
"An excellent book that provides an in-depth phenomenological and philosophical study of memory." -Choice

". . . a stunning revelation of the pervasiveness of memory in our lives." —Contemporary Psychology

Studies in Continental Thought
384 pp
cloth $49.95 paper $22.95

Now in a new edition!

IMAGINING



A Phenomenological Study, Second Edition

Edward S. Casey

"This scrupulous, lucid study is destined to become a touchstone for all future writings on imagination."
—Library Journal

Studies in Continental Thought
256 pp
cloth $39.95 paper $18.95

CHARLES S. PEIRCE'S PHILOSOPHY OF SIGNS

Essays in Comparative Semiotics

Gérard Deledalle

"This is essential reading for anyone who wants to find common ground between the best of American semiotics and better-known European theories. Deledalle has done more than anyone else to introduce Peirce to European audiences, and now he sends Peirce home with some new flare."
—Nathan Houser, Director, Peirce Edition Project

December 2000 304 pp
cloth $39 95

at bookstores

INDIANA UNIVERSITY PRESS

Orders: 1-800-842-6796
iupress.indiana.edu

Critical Thinking

Nietzsche and the Philology of the Future

JAMES I. PORTER

Drawing on Nietzsche's prolific early notebooks and correspondence, this book challenges the polarized picture of Nietzsche as a philosopher who abandoned classical philology. By showing how frequently the "later" Nietzsche appears in the early writings, the author hopes to provoke reflection on the adequacy of the developmental logic that has been a controlling factor in Nietzsche's reception.

$19.95 paper $60.00 cloth

NOW IN PAPERBACK

The Complete Works of Friedrich Nietzsche

Volume 3: Human, All Too Human (I)

FRIEDRICH NIETZSCHE

Translated, with an Afterword, by Gary Handwerk

This volume is the earliest of Nietzsche's works in which his philosophical concerns and methodologies can be glimpsed. It is part of an edition that has received universal praise: "It has revolutionized our understanding of one of the greatest German thinkers"; "Scholars can be confident for the first time of having a trustworthy text."

$14.95 paper $29.95 cloth

The Instant of My Death /*Demeure*: Fiction and Testimony

MAURICE BLANCHOT / JACQUES DERRIDA

Translated by Elizabeth Rottenberg

This volume, a powerful short prose piece by Blanchot with an extended essay by Derrida, records a remarkable encounter in critical and philosophical thinking.

Meridian: Crossing Aesthetics

$12.95 paper $29.50 cloth

Art as a Social System

NIKLAS LUHMANN

Translated by Eva M. Knodt

This is the definitive analysis of art as a social and perceptual system by Germany's leading social theorist of the late 20th century. It combines three decades of research in the social sciences, phenomenology, evolutionary biology, cybernetics, and information theory with an intimate knowledge of art history, literature, aesthetics, and contemporary literary theory.

Meridian: Crossing Aesthetics

$24.95 paper $65.00 cloth

The Invention of Dionysus

JAMES I. PORTER

This book argues that *The Birth of Tragedy*, Nietzsche's first book, does not mark a rupture with his prior philosophical undertakings but is, in fact, continuous with them and with his later writings as well. It shows that many of the book's elements are reminiscent of Nietzsche's earlier revisions of philology and anticipate the later writings.

$16.95 paper $45.00 cloth

Problems of Moral Philosophy

THEODOR W. ADORNO

Edited by Thomas Schröder, Translated by Rodney Livingstone

This volume makes Adorno's lectures on the problems of moral philosophy available for the first time to English-speaking readers.

$39.50 cloth

Introduction to Sociology

THEODOR W. ADORNO

Edited by Christoph Gödde, Translated by Edmund Jephcott

This book distills decades of distinguished work in sociology by one of the 20th century's most influential thinkers.

$39.50 cloth

Pascalian Meditations

PIERRE BOURDIEU
Translated by
Richard Nice

A brilliant example of Bourdieu's unique ability to link sociological theory, historical information, and philosophical thought.

$18.95 paper $49.50 cloth

Deconstruction as Analytic Philosophy

SAMUEL C. WHEELER III

Twelve essays treat the thought of "deconstructive" philosophers from the perspective of analytic philosophy and relate such thinkers as Davidson, Quine, and Wittgenstein to Derrida and de Man.

Cultural Memory in the Present

$19.95 paper $49.50 cloth

Belief

GIANNI VATTIMO
Translated by
Luca D'Isanto and
David Webb

In this highly personal book, one of Europe's foremost contemporary philosophers confronts the theme of faith and religion. He argues that there is a substantial link between the history of Christian revelation and the history of nihilism, in particular as the latter appears in the work of Nietzsche and Heidegger, Vattimo's philosophical specialty.

Cultural Memory in the Present

$12.95 paper $39.50 cloth

The Words of Selves
Identification, Solidarity, Irony

DENISE RILEY

In this extended meditation on the language of the self within contemporary social politics, the author ponders the question: What does it matter what you say about yourself? She studies why the requirement *to be* a something-or-other should be so hard to satisfy in a manner that rings true in the ears of its own subject.

Atopia: Philosophy, Political Theory, Aesthetics

$18.95 paper $49.50 cloth

The Confession of Augustine

JEAN-FRANÇOIS LYOTARD
Translated by
Richard Beardsworth

This remarkable posthumous work by one of the leading philosophers of the 20th century engages Augustine's *Confessions*, one of the major canonical works of world literature and the very paradigm of autobiography as a definable genre of writing. Lyotard approaches his subject by returning to his earliest phenomenological training.

Cultural Memory in the Present

$12 95 paper $39.50 cloth

World Spectators

KAJA SILVERMAN

"This original and important book demonstrates the inseparability of philosophy and psychoanalysis for any serious attempt to answer a question so profoundly relevant to the very nature of our being that it does not 'belong' to any one discipline: the question, as Silverman puts it, of what it means for the world that each one of us is in it. The book has a remarkable clarity; Silverman makes the most complex argument seem like a perfectly natural, and absolutely necessary, movement of thought."
—Leo Bersani,
University of California,
Berkeley

Cultural Memory in the Present

$16.95 paper $45.00 cloth

The Spirit of Utopia

ERNST BLOCH
Translated by
Anthony A. Nassar

"When this book was first published, it had a profound effect on major thinkers and artists in Weimar Germany. . . . Bloch's book laid the groundwork for thinkers like Adorno and Benjamin."
—Jack Zipes,
University of Minnesota

Meridian: Crossing Aesthetics

$19.95 paper $55.00 cloth

Stanford
University Press

800·872·7423 | www·sup·org

Forthcoming!

HEGEL'S DIALECTICAL LOGIC

ERMANNO BENCIVENGA

This clear, accessible account of Hegelian logic offers comprehensive coverage of Hegel's important works. Bencivenga makes a case for Hegelian logic's enormous seductiveness, its surprising presence in the collective consciousness, and the dangers associated therewith.
October 2000 176 pp. $35.00

OF MINDS AND MOLECULES

New Philosophical Perspectives on Chemistry

Edited by NALINI BHUSHAN and STUART ROSENFELD

This is the first anthology devoted exclusively to work in the philosophy of chemistry. The essays adopt distinctive philosophical perspectives on chemistry and collectively offer both a conceptualization of and a justification for this emerging field.
2000 320 pp. $55.00

FAITH WITH REASON

PAUL HELM

Helm investigates what religious faith is and what makes it reasonable. Helm argues that the reasonableness of faith depends not only on beliefs about the world but also on beliefs about oneself and on what one is willing to trust.
2000 196 pp. $39.95

HUME'S ABJECT FAILURE

The Argument Against Miracles

JOHN EARMAN

This vital study offers a new interpretation of Hume's famous "Of Miracles," which notoriously argues against the possibility of miracles. By situating Hume's argument in the context of the eighteenth-century debate on miracles, Earman shows Hume's argument to be largely unoriginal and chiefly without merit where it is original.
2000 256 pp. paper $24.95 cloth $49.95

New in paperback!

TIME, TENSE, AND CAUSATION

MICHAEL TOOLEY

"Tooley's provocative book makes an extremely valuable contribution to the literature on the philosophy of time and is mandatory reading for anyone working in the area."—*Mind.*
1997 (paper 2000) 416 pp.
paper $24.95 cloth $45.00

THE ANCIENT QUARREL BETWEEN PHILOSOPHY AND POETRY REVISITED

Plato and the Greek Literary Tradition

SUSAN B. LEVIN

In this study, Levin argues that Plato takes the Greek literary tradition as a key locus of ideas with respect to linguistic issues in the *Cratylus*, *Phaedo*, and the *Republic.*
2000 208 pp. $39.95

New in paperback!

NATURALISM IN MATHEMATICS

PENELOPE MADDY

"[Maddy] writes in a clear, forthright and challenging style. Her book is eminently readable, instructive, and thought-provoking."—*The Journal of Symbolic Logic.*
1998 (paper 2000) 264 pp.
paper $19.95 cloth $45.00

THE CHARACTER OF LOGIC IN INDIA

BIMAL KRISHNA MATILAL

Edited by JONARDON GANERI and HEERAMAN TIWARI

In this book, Matilal brings together two decades of thought about the origins, development, and nature of Indian logical theory. His distinctive treatment of the philosophical issues allows our understanding of classical Indian writers to be enriched by concepts drawn from contemporary analytical theory.
2000 192 pp. $12.95

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our website at **www.oup.com**

New in paperback!

POINTS OF VIEW

A. W. MOORE

Moore argues that it is possible to think about the world from no point of view. He explores issues such as the nature of persons, the subject-matter of mathematics, realism and anti-realism, value, and God. The result is a powerful critique of our own finitude.

1997 (paper 2000) 328 pp.
paper $19.95 cloth $35.00

New in paperback!

SEX AND SOCIAL JUSTICE

MARTHA C. NUSSBAUM

"*Sex and Social Justice* is highly readable, and very engaging. It is elegantly written and carefully argued."
—Alan Ryan, *The New York Times Book Review.*

1999 (paper 2000)
448 pp. paper $15.95
cloth $35.00

THEMES IN HUME

The Self, the Will, Religion

TERENCE PENELHUM

Since the 1950s, Penelhum has been a leading contributor to studies on the thought of David Hume. In this collection, he presents a selection of the best of his essays on Hume.

2000 296 pp. $55.00

GOD AND GOODNESS

HUGH RICE

Rice explains why belief in God need not be seen as a strange or irrational belief, but can be a natural extension of our ordinary ways of thinking. He suggests that we should think of God in an abstract way, and he offers a satisfying account of the relationship between God and goodness.

2000 156 pp. $29.95

OXFORD STUDIES IN ANCIENT PHILOSOPHY

Volume XVIII

Edited by DAVID SEDLEY

This is an annual volume of original articles on all aspects of ancient philosophy. This year's contributions range over a thousand years of philosophy, from the Presocratics to Boethius. Particularly prominent in the volume are Aristotle and the Stoics.

2000 256 pp. paper $35.00 cloth $60.00

New in paperback!

SEXTUS EMPIRICUS: AGAINST THE ETHICISTS

(Adversus Mathematicos XI)

Translated with an introduction and commentary by RICHARD BETT

"Bett's careful translation...ought to be the main English translation consulted....Judicious and learned... Bett's work is a distinguished addition to the literature on skepticism."—*Bryn Mawr Classical Review.*

(Clarendon Later Ancient Philosophers)

1997 (paper 2000) 336 pp.
paper $24.95 cloth $65.00

EMOTION AND PEACE OF MIND

From Stoic Agitation to Christian Temptation

RICHARD SORABJI

Sorabji presents a ground-breaking study of ancient Greek views of the emotions and their influence on subsequent theories and attitudes. Focussing on the Stoics, Sorabji also draws on a vast range of texts to give a rich historical survey of how Western thinking about the emotions developed.

2000 440 pp. $45.00

TRUTH, RATIONALITY, AND PRAGMATISM

Themes from Peirce

CHRISTOPHER HOOKWAY

Hookway presents a series of studies of themes from the work of Charles S. Peirce (1839–1913). These themes center on the question of how we are to investigate the world rationally. Hookway shows how Peirce's ideas about this continue to play an important role in contemporary philosophy.

2000 280 pp. $45.00

Prices are subject to change and apply only in the US.
To order, or for more information, please call
1-800-451-7556. In Canada, call 1-800-387-8020.
Visit our website at **www.oup.com**

OXFORD
UNIVERSITY PRESS

Outstanding Scholarship from Cambridge

The Path of the Law and its Influence

The Legacy of Oliver Wendell Holmes

***Steven J. Burton,** Editor*

Cambridge Studies in Philosophy and Law

0-521-63006-1 Hardback $64.95

Philosophy of Logics

Susan Haack

0-521-21988-4 Hardback $47.50
0-521-29329-4 Paperback $24.95

Berkeley's *Principles and Dialogues*

Background Source Materials

***C. J. McCracken** and*
***I. C. Tipton,** Editors*

Cambridge Philosophical Texts in Context

0-521-49681-0 Hardback $59.95
0-521-49806-6 Paperback $22.95

Cambridge Companion to Malebranche

***Steven Nadler,** Editor*

Cambridge Companions to Philosophy

0-521-62212-3 Hardback $54.95
0-521-62729-X Paperback $19.95

The Right to Privacy

Ellen Frankel Paul,
***Fred D. Miller, Jr.,** and*
***Jeffrey Paul,** Editors*

Social Philosophy and Policy

0-521-78621-5 Paperback $24.95

Reading Neoplatonism

Non-discursive Thinking in the Texts of Plotinus, Proclus, and Damascius

Sara Rappe

0-521-65158-1 Hardback $59.95

The Reception of Kant's Critical Philosophy

Fichte, Schelling and Hegel

***Sally Sedgwick,** Editor*

0-521-77237-0 Hardback $59.95

Now in paperback...

Lectures on Ethics

Immanuel Kant

***Peter Heath** and*
***J. B. Schneewind,** Editors*

Cambridge Philosophical Texts in Context

0-521-78804-8 Paperback $24.95

Aesthethics and Ethics

Essays at the Intersection

***Jerrold Levinson,** Editor*

Cambridge Studies in Philosophy and the Arts

0-521-78805-6 Paperback $19.95

Philosophical Meditations on Zen Buddhism

Dale S. Wright

Cambridge Studies in Religious Traditions

0-521-78984-2 Paperback $19.95

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

40 West 20th Street,
New York, NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cup.org
AmEx/MasterCard/VISA accepted.
Prices subject to change.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NUMBER 9
SEPTEMBER 2000

page
477 *Consequential Evaluation and Practical Reason* Amartya Sen
503 *Two Dogmas of Belief Revision* Hans Rott

BOOK REVIEWS
523 Wesley C. Salmon: *Causality and Explanation* Paul Humphreys

528 NEW BOOKS: ANTHOLOGIES

CALCUTTA UNIVERSITY LIBRARY

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVII, NO. 9, SEPTEMBER 2000

CONSEQUENTIAL EVALUATION AND PRACTICAL REASON*

The subject of this essay is the relevance and reach of consequential evaluation in practical reason. A good starting point for this analysis, I argue, is the need to take responsibility for the consequences of one's choice. The demands of responsibility of the person making a choice (of actions, strategies, or other decision variables in practical reason) relate both to (1) the discipline of *evaluation*, and to (2) the discipline of *choice* based on that evaluation. Because of these disciplines, consequential evaluation can, I argue, systematically combine very diverse concerns, including taking responsibility for the nature of one's actions (and related considerations that have figured prominently in the deontological literature), without neglecting other types of consequences (on which some of the narrower versions of consequential reasoning—such as utilitarianism—have tended to concentrate). There are advantages, I argue, in a broad but integrated framework of this kind.

This paper is about consequential evaluation seen as the discipline of responsible choice based on the chooser's evaluation of states of affairs, including consideration of all the relevant consequences viewed in the light of the exact circumstances of that choice. It is not about how the commonly used *term* 'consequentialism' should be used. Whether consequential evaluation, as explored here, should be called by the name 'consequentialism' or not is a subsidiary and

---

* The essay draws on my Rosenthal Lectures at the Northwestern University Law School in Chicago, given in September 1998, under the title "The Domain of Justice." For helpful comments, I am grateful to Ron Allen, Sudhir Anand, Elizabeth Anderson, Christian Barry, Akeel Bilgrami, G.A. Cohen, Susan Hurley, Isaac Levi, Martha Nussbaum, Onora O'Neill, Philip Pettit, and Thomas Scanlon, and to the participants in seminar discussions at the Northwestern Law School, and at Cambridge, Oxford, and Columbia University.

0022-362X/00/9709/477–502

rather uninteresting issue. The fact is that the term 'consequentialism' was devised by enemies rather than proponents of consequential evaluation, and it has been invoked mainly to be refuted, often with colorful counterexamples that have added some spice—and much fun—in moral philosophy. To admit to being a "consequentialist" is almost like introducing oneself as a "wog" or a "frog" or a "limey," which can be instantly delineating without being descriptively rich. Indeed, the term is unattractive enough (even if it did not remind one of diseases like botulism) to be sensibly bequeathed to anyone who wants to take it away.

As it happens, however, my characterization of consequential evaluation is not in conflict with at least one standard definition of consequentialism as the requirement that any choice of actions (or rules, or strategies, or whatever) be based on selecting an alternative that produces no worse an overall outcome than any other available alternative. For example, it is entirely compatible with Philip Pettit's definition of consequentialism, as presented in the introduction to his distinguished collection of essays on this subject, called *Consequentialism*.[1] "Roughly speaking," Pettit says, "consequentialism is the theory that the way to tell whether a particular choice is the right choice for an agent to have made is to look at the relevant consequences of the decision; to look at the relevant effects of the decision on the world" (*ibid.*, p. xiii). His definition also has the merit of linking the exercise of choice immediately to the idea of taking responsibility for the consequences of one's choice, and is thus close to the approach explored in this essay.

The discipline of this responsibility has not always been adequately recognized within consequentialist ethics. The standard versions of utilitarian ethics have been, I would argue, especially short in this, in particular through ignoring all consequences other than utilities, even when they are part and parcel of the state of affairs (for example, the actions of particular agents that have actually occurred). This has followed from the utilitarian program of combining consequentialism with *additional* demands, particularly "welfarism," which insists that states of affairs must be judged exclusively by the utility information (such as happiness or desire fulfillment) related to the respective states—no matter what the other features of the consequent state of affairs may be, such as the performance of particular acts (however nasty), or the violation of other people's liberties (however personal).[2]

---

[1] Aldershot: Dartmouth, 1993.

Also, "critics" of consequential evaluation often do not take sufficient note of the reach of consequential reasoning shorn of the arbitrary limits imposed by such additional restrictions as welfarism. Indeed, since these additional constraints also reduce the opportunity of taking adequate note of the responsibilities of consequential evaluation, a critic of consequential evaluation who assumes that these constraints must be a part of that evaluation is led to do far less than justice to the potential reach and extensive relevance of consequential reasoning.

How, then, does broad consequential evaluation compare with deontological approaches? This comparison is difficult to make because deontology can take very many different forms. Some deontological approaches have very limited reach because of their insistence on making *consequence-independent* judgments (I shall presently consider a classic argument of this kind). Others allow particular consequences to be considered, but rule out the possibility of some or all *trade-offs* in taking note of diverse concerns (libertarian deontology is a prominent example of this and imposes various nonrelaxable constraints).[3] In contrast, some permissive deontologists seem to admit many different kinds of concerns, if only in stages.[4] Trade-offs seem to be allowed, in effect, when the different conflicting concerns are balanced against each other in broad deontological analysis, even if the overall discipline of these broad-based judgments remains somewhat veiled and unexplicit.[5]

---

[2] Indeed, utilitarian reasoning is an amalgam of three distinct axioms: (1) consequentialism, (2) welfarism, and (3) sum ranking (the last stands for the requirement that utilities of different people must simply be added up to assess the state of affairs, paying no attention to, say, inequalities). On the factorization of utilitarianism, see my "Utilitarianism and Welfarism," this JOURNAL, LXXVI, 9 (September 1979): 463-89; and Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (New York: Cambridge, 1982)—see particularly our "Introduction."

[3] The most elegant exposition of the libertarian approach can be found in Robert Nozick, *Anarchy, State and Utopia* (New York: Basic, 1974).

[4] Well illustrated by the variety of arguments on different sides that get sympathetic hearing in Williams's broad deontological analysis; see, for example, his "A Critique of Utilitarianism," in J.J.C. Smart and Williams, *Utilitarianism: For and Against* (New York: Cambridge, 1973). See also Thomas Nagel, *Mortal Questions* (New York: Cambridge, 1979); Derek Parfit, *Reasons and Persons* (New York: Oxford, 1984); Thomas Scanlon, *What We Owe to Each Other* (Cambridge: Harvard, 1999); among other contributions.

[5] I have argued elsewhere that the substantive gap between some versions of broad deontology and broad consequentialism may not be very great; see my "Evaluator Relativity and Consequential Evaluation," *Philosophy and Public Affairs*,

It is, thus, difficult to give one simple answer to the question that may well be asked: 'Why use broad consequential evaluation, rather than deontology?' The contrast depends on what kind of deontology is being considered in the comparison. In contrast with *consequence-independent* deontology, or *trade-off-barred* deontology, broad consequential evaluation has considerably more reach and range, in being able to accommodate diverse moral concerns that have claims to our attention. In comparison with permissive kinds of deontology, broad consequential evaluation can claim to have, at least, a more explicit—and somewhat more integrated—framework of judgmental evaluation.

There is, in fact, a further argument, which I shall not explore in any detail, but to which I shall point (since it is, I believe, of some importance in seeing the advantages of broad consequential reasoning). By linking ideas of "goodness" and "rightness" in a general and consolidated framework, an adequately broad consequential framework avoids the kind of alienation that seems to be involved in such statements (frequently invoked, in one form or another, in high deontology): "This would be ethically the best thing to do; however for ethical reasons, I must not do it."[6] Ethics is complicated enough because of our different moral beliefs and the external disagreements we all tend to face, and it is unfortunate to have to add to all this some signs of internal tension as well.

I. CONSEQUENTIAL EVALUATION AND CONSEQUENCE-INDEPENDENT DEONTOLOGY

I begin, however, not with contemporary debates, but one that is reported to have occurred more than two millennia ago. In the Indian epic, the *Mahabharata*, the composition of which is usually attributed to the period around the second or third century before Christ, there is an interesting conversation that occurs between Arjuna, the great warrior hero of the epic, and Krishna, his friend and advisor, on the eve of the massive battle at Kurukshetra (not far from the city of Delhi). The battle is between the Pandavas, the virtuous royal family presided over by Yudhisthira, Arjuna's eldest brother and legitimate heir to the throne, and the Kauravas, their cousins, who have wrongly usurped the kingdom. Nearly all the other royal fami-

---

xii, 2 (Spring 1983): 113-32. It is partly a question of what are made explicit and what are weighed into the judgment in somewhat implicit ways.

[6] For related issues, see Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs*, xiii, 2 (Spring 1984): 14-171; and Samuel Scheffler, "Agent-Centred Restriction, Rationality, and the Virtues," *Mind*, xciv, 375 (1985): 409-19; both reprinted in Scheffler, *Consequentialism and Its Critics* (New York: Oxford, 1988), pp. 93-133, 243-60.

lies in the Indo-Gangetic plain have joined one side or the other, and the two armies include a considerable proportion of able-bodied men in the land.

As Arjuna and Krishna see the armies on the two sides and reflect on the gigantic battle that is about to begin, Arjuna asks whether all this is worth it. He does not doubt that theirs is the right cause, and that this is a just war, and also that his side will definitely win the battle given its relative strength—not least because of Arjuna's own remarkable skills as a warrior and a general. But, Arjuna observes, so many people will die in this battle. He also recognizes that he himself will have to kill masses of people. Arjuna also notes that many people who will be killed, on both sides, are persons for whom he has affection, and many of them have done nothing that is particularly disreputable other than agreeing (often out of kinship loyalties or other ties) to back one side or the other in this great subcontinental battle.

Arjuna tells Krishna that he does not want to fight, and that maybe they should simply let the unjust Kauravas rule the kingdom they have usurped, which may be the lesser of the two evils. Krishna argues against this, and his response takes the form of articulating principles of action that have been repeated again and again in Indian moral philosophy. Indeed, with Krishna's gradual transformation from being a noble but partisan patron of the Pandavas in the epic to being an incarnation of God, as he is in later Hinduism, the conversation with Arjuna has become a document of great theological importance, called the *Bhagavadgeeta*, or the *Geeta* for short.

Krishna points to Arjuna's duty to fight, irrespective of his evaluation of the consequences. It is a just cause, and as a warrior and a general on whom his side must rely, he cannot waver from his obligations (no matter what results from that). Krishna's high deontology has been deeply influential in moral debates in the subsequent millennia. It is, I suppose, a tribute to the power of pure theory that even Mahatma Gandhi—no less—felt deeply inspired by Krishna's words on doing one's duty irrespective of consequences, even though the duty in this case was for Arjuna to fight a violent war (not in general a cause to which Gandhi could be expected to warm).

Krishna's moral position is also eloquently endorsed by T.S. Eliot, in a poem, "The Dry Salvages," in the *Four Quartets*.[7] Eliot summarizes Krishna's view in the form of an admonishment: "And do not think of the fruit of action./ Fare forward." Eliot explains: "Not fare well,/ But fare forward, voyagers" (*ibid.*, p. 31). I intend here to take the other side—that of Arjuna—and proceed from the basic idea that one must take responsibility for the consequences of one's actions and choices, and that this responsibility cannot be obliterated by any pointer to a consequence-independent duty or obligation. "Fare," I would argue, "well," and not just "forward." I should add, just as an incidental matter, that even though as a religious document the *Geeta* is interpreted to be firmly on Krishna's side (it is alleged that he "won" the debate), the epic *Mahabharata* in which the conversation originally occurs gives both sides much room to develop their respective arguments, with the evident presumption that this is an argument in which there are two reasonable sides.

---

[7] London: Faber and Faber, 1944.

II. GENERAL CONSEQUENTIALISM AND PARTICULAR CONSEQUENTIALIST SYSTEMS

Consequential evaluation can be analyzed at two different levels. First, we can study the general demands of basing evaluation and choice on the assessment of consequences and, at this common level, the discipline of consequential evaluation leaves room for different ways of getting to a fuller consequentialist theory. Second, this general analysis can be combined with the exploration of a specific evaluative theory, or a class of evaluative theories that share the important feature of focusing on responsibility for choice. I am concerned here with both these exercises.

In the general analysis, it is especially important to recognize that the very specific demands associated with some of the more famous consequentialist systems, such as welfarism (on which utilitarian ethical theory relies), are not *required* by the discipline of consequentialism itself. It may be, arguably, permissible to combine these additional demands with consequential evaluation (as utilitarianism does).[8] But they are certainly not mandatory for consequential evaluation in general. I have commented on this subject already. But now, going beyond this, I must also discuss what kind of further specification may be sensibly pursued to get a fuller consequential theory (different from utilitarianism)—a theory or an approach that would be particularly in line with the basic motivation of taking responsibility for one's choice. To that exercise I shall now turn, beginning first with identifying particular areas in which further specification—additional to the general demands of consequential evaluation—would have to be made.

---

[8] I say 'arguably' here, because it can also be argued, on the other side, that by confining attention to utility information only, welfarism goes against the spirit of taking responsibility for *all* the consequences of one's choice. I do not, however, intend to insist, on this ground, that we exclude utilitarianism from the class of genuinely consequentialist theories altogether. It is, rather, my claim that utilitarianism is deeply flawed as a substantive consequentialist theory, and this is where its tension with the basic motivation of taking responsibility for choice comes in. The problem lies, on this reading, in its inadequacy, rather than in its inadmissibility.

There is a need, I would argue, for more articulation on at least three different subjects, related to the discipline of consequential evaluation.

(1) *Situated evaluation.* From which perspective should the evaluation be done? The argument from responsibility would indicate that the person making the choice cannot escape the necessity to take note of her own position vis-à-vis the actions and their consequences. This requirement of "situated evaluation" may or may not be combined with the demand (made familiar by utilitarian ethics) that the evaluation from *every* situated position must somehow be exactly *the same.* This is an "invariance requirement" for which an additional argument would be needed, if it were to be proposed.[9] But I do not intend to propose it.

(2) *Maximizing framework.* As Samuel Scheffler[10] rightly points out, "*maximizing* rationality" is central to consequential evaluation. As he puts it: "The core of this conception of rationality is the idea that if one accepts the desirability of a certain goal being achieved, and if one has a choice between two options, one of which is certain to accomplish this goal better than the other, then it is, *ceteris paribus,* rational to choose the former over the latter" (*ibid.,* p. 252). I shall have to talk more, later on, about the extensive implications of this principle, but one immediate question concerns whether the maximizing framework can be used even when the options cannot be completely ranked vis-à-vis each other. What if the evaluation cannot sensibly go beyond, say, a partial ordering? As would be readily seen, Scheffler's statement of the maximizing principle does not demand completeness of ranking. But is his silence on completeness an omission, or is his characterization of maximization adequate as it stands? I would argue that there is no omission here, but I must go further into this issue (more on this presently). I should state here, however, that not to demand completeness does not exclude the possibility that the evaluative ordering may, in fact, be complete. If completeness were to be specifically demanded (rather than only permitted), however, then some justification for this demand would be needed. But I make no such demand.

---

[9] On the nature and use of "invariance requirements," see my *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970; republished: Amsterdam: North-Holland, 1979), especially chapters 7 and 8; and Kevin Roberts, "Possibility Theorems with Interpersonally Comparable Welfare Levels," *Review of Economic Studies,* XLVII, 2: 409-20.

[10] "Agent-Centred Restriction, Rationality, and the Virtues" (1985), reprinted in *Consequentialism and Its Critics.*

(3) *Nonexclusion of state components.* Are all features of the consequent states of affairs potentially relevant? In terms of the discipline of consequential evaluation, it would be arbitrary to exclude a priori any particular component of the state as being beyond the pale of consideration. This requirement of "nonexclusion of state components" does not, by itself, rule out the possibility of insisting, as is done in utilitarian ethics, that *only utilities* (in the form of, say, happiness or desire fulfillment) in the states of affairs must ultimately be taken into account in ethical evaluation. If that substantive restriction were to be imposed, then a justificatory defense of it would have been needed (since it is not entailed by the discipline of consequential evaluation in general). But I impose no such restriction.

I shall now briefly outline the broad characteristics of a specific *class* of consequentialist approaches (within the limits of the chosen requirements), which have, I would argue, some claim to our attention.

III. RESPONSIBILITY AND SITUATED EVALUATION

From which perspective should the evaluation be done? The demand of situated evaluation requires that a person not ignore the particular position from which she is making the choice. Consider, for example, the parent of a child for whom she is choosing a particular baby food. The requirement of situated evaluation does not, in any way, vindicate smugness about one's contingent level of ignorance, and does not deny the need for the person to find out, if reasonably possible, more about what others know or see (for example, that the baby food with which the parent is familiar might have been shown to be harmful). Nor does it deny the relevance of broader sympathies (for example, a parent may well ask whether it is right that her child should have the benefit of some baby food to which other children do not have access). What is denied is the possibility of ignoring the person's own responsibilities in her particular situation, in this case that of being a parent of this child.

Even if the decision is to respond to some considerations coming from outside her situation (such as not using a baby food of which the neighbors are suspicious, or not giving one's child a baby food that the neighbors cannot afford to give to their children), that decision would have to be worked out from—and in this sense be compatible with—the responsibility she has toward her child for whom she is choosing a baby food (the "other" considerations and their importance would have to be evaluated by the parent *situated in her own position*). She has to take responsibility for not only her choices of substantive actions (going beyond evaluation), but also for her evaluations on which the choices are to be based. She has to do the evaluation, taking note—as Arjuna does—of her own position in the events and of her responsibility for things that may happen as a result of what she does.

This valuational issue raises a question of great importance regarding the relationship between the person who is doing the evaluation (as well as the choosing) and the particular circumstances in which she is situated. Arjuna is bothered not merely by the fact that many will die if the war were to take place, but also by the fact that he will himself be killing lots of people and by the further fact that many of the people to be killed are persons for whom he himself has affection. These are the things that contribute to the bad nature of the events that would occur due to the war, seen particularly from the position of Arjuna himself. This realization is part and parcel of his taking responsibility for his choices and their consequences. Another observer who is uninvolved in these events need not attach any special importance to the fact that Arjuna (not *he*, but *Arjuna*) will be killing people, and that among the dead will be people for whom Arjuna (not *he*, but *Arjuna*) feels closeness and affection. Arjuna cannot reasonably take a similarly detached view of the consequences of his choice, since he is directly involved in making this choice.

The evaluations can, thus, be reasonably situated in the life of the person doing the evaluation and making the choices. This contrasts quite sharply with the utilitarian formula that the evaluation must be, in every way, independent of the evaluator and, in particular, must take the very specific form of maximizing the sum total of utilities. Such has been the hold of utilitarianism within the general discipline of consequential evaluation that rather than taking the utilitarian claim of evaluator-independence as a very special case (at best admissible and certainly not mandatory), consequentialism is often defined as having to depend inescapably on evaluator-independent valuations.

I should also comment on the fact that the discipline of position-related evaluator-relativity, the fuller implications of which I have tried to explore elsewhere,[11] must be distinguished from a claim that anyone is free to evaluate consequences as he likes. Far from being free to evaluate consequences as he likes, Arjuna has the responsibility—indeed *cannot escape* the responsibility—to take note of the special badness of the events as he must evaluate them, because he himself would have to do some of the killing, and victims would often be quite innocent persons for whom Arjuna has affection. It is not so much a *license* to evaluate the outcomes as one likes, but a *requirement* that one must take note of the contingent connections and circumstances that characterize Arjuna's situation in this choice problem. Responsibility—and this is the important point here—demands situated valuation by the agents.

---

[11] "Rights and Agency," *Philosophy and Public Affairs*, XI, 1 (Winter 1982): 3-39, and "Well-being, Agency and Freedom: The Dewey Lectures 1984," this JOURNAL, LXXXII, 4 (April 1985): 169-21.

It is also important to see that evaluator-relativity need not violate any requirement of "impersonality" that may be imposed on the discipline of ethics. Indeed, impersonality must not be confused with what may be called *impositionality*—having to ignore the relevance of a chooser's position or situation vis-à-vis the choices and results. In contrast with impositionality, position-relative impersonality requires, to put it in mildly mathematical terms, that parametric note be taken of the respective *positions* of the different persons, but not of the exact personal *identities* involved.[12]

IV. INCOMPLETENESS AND MAXIMIZATION

I turn now to the second issue, the possibility of incompleteness of ranking in consequential evaluation. It is sometimes presumed that a consequentialist approach must require that every possible state of affairs be comparable—and be clearly ranked—vis-à-vis every other. This presumed requirement has sometimes been seen, with justice, as unlikely to be fulfilled, and it has been seen as a source of problems for consequentialist ethics. How can we always compare every alternative with every other? Can we always find a best alternative?

Maximization does not, in fact, demand that all alternatives be comparable, and does not even require that a best alternative be identifiable. It only requires that we do not choose an alternative that is worse than another that can be chosen instead (as Scheffler (*op. cit.*, p. 252) has rightly noted). There is some confusion on this technical point since in common parlance, the term 'maximization' is used in many different ways. Sometimes the term is used to indicate that we must choose a *best* alternative. This is, technically, better described as *optimization* (also a concept much used in economics and decision theory). The basic definition of maximization in the foundational literature on set theory and relational analysis is that of selecting an alternative to which there is none better. Maximization and optimization coincide when the ordering is *complete*, which it may or may not be.[13] Indeed, for optimization to work there should *exist* a best alternative to be chosen (not necessarily a

---

[12] On this, see my "Positional Objectivity," *Philosophy and Public Affairs*, XXII, 2 (Spring 1993): 126-45.

[13] The exact relations between optimization and maximization, and the conditions that make them coincide with each other, are identified in my "Maximization and the Act of Choice," *Econometrica*, LXV, 4 (July 1997): 745-79.

uniquely best alternative, but a best alternative nevertheless). Maximization does not require that.

For example, if there happens to be some incompleteness so that two options A and B cannot be ranked vis-à-vis each other, but each of them is better than all the other alternatives, then maximization would require only that one of those two be chosen (both are maximal). This is the way maximality is defined in the mathematical literature, both in pure set theory (for example, by Nicolas Bourbaki), and in axiomatic economic analysis (for example, by Gerard Debreu).[14] I have discussed elsewhere[15] this distinction and its far-reaching implications.

The distinction can be a matter of great practical importance. Take the old story of Buridan's ass. The hapless donkey saw two haystacks that it could not rank vis-à-vis each other.[16] Since donkeys are presumed to be full-blooded optimizers rather than relaxed maximizers, Buridan's ass could not choose either haystack (since neither was shown to be at least as good as the other), and it thus died of starvation. It starved since it could not rank the two haystacks (there was no "best" or "optimal" choice), but of course each would have generated a better consequence than starvation. Even if the donkey failed to rank the two haystacks, it would have made sense—good consequentialist sense—for it to choose either rather than neither. Consequential evaluation needs maximization, but not optimization.

V. STATES, ACTIONS, MOTIVES, AND PROCESSES

I turn now to the third issue—what to include in the description of states of affairs, in using consequentialist logic. There is no general reason for an a priori exclusion of any part of the state of affairs in evaluating that state of affairs. Some ethical theories (like utilitarian ethics) insist, however, that nonutility features, such as actions, must not be considered to be of any value or disvalue on their own (rather, only for the utilities or disutilities they generate). But since there are good reasons to be concerned about some of these features, such as actions, motives, and the like (for reasons which are close enough to those which move deontologists, among others, to take note of them, in their own special way), the utilitarian exclusion ends up being an arbitrary exclusion of a class of reasoned demands. I have discussed this issue elsewhere,[17] and shall not pursue it further here. In terms of the example considered earlier, I should note that Arjuna's argument against killing people in general, and killing friends and relations in particular, has clear implications for his choice of action.

---

[14] Bourbaki, *Élements de Mathématique* (Paris: Hermann, 1939) —*Volume I, Theory of Sets* (Reading, MA: Addison-Wesley, 1968); Debreu, *Theory of Value* (New York: Wiley, 1959; republished, New Haven: Yale, 1971). Isaac Levi—*Hard Choices: Decision Making under Unresolved Conflict* (New York: Cambridge, 1986)—has explored a different way of basing choices on incomplete rankings, which I shall not pursue here.

[15] *Collective Choice and Social Welfare*, chapter 1; and "Maximization and the Act of Choice."

[16] There is a less interesting version of the story of Buridan's ass, according to which it was *indifferent* between the two hay stacks and could not decide which to choose. But if it were really indifferent, then either haystack would, clearly, have been as good as the other, and even an optimizing donkey need not have faced an impasse.

Indeed, the idea of responsibility for the consequences of one's actions can hardly be compatible with ignoring some of the consequences one takes seriously—and has reason to take seriously. This issue relates closely to the tension on which Scheffler concentrates in arguing that what "makes so much trouble for views that incorporate agent-centred restrictions, is what we may call *maximizing* rationality" (*op. cit.*, p. 252). Agent-centered concerns need not, of course, be presented in the form of constraints or "restrictions" (as they are in some deontological theories), but agent-specific sensitivity can be incorporated within the maximizing framework of consequential reasoning by taking evaluative note of the actions of agents, which are part of the state of affairs. The demands of responsibility certainly call for agent-sensitive evaluation. Arjuna cannot treat his own killing of people (and even other deaths that would occur if he were to allow the war to go on, with his support) in the same way that he can treat deaths that result from other causes or other agencies. There need not really be any tension between maximizing rationality and agent-sensitive reasoning (central for consequential responsibility *as well as* for deontology), unless parts of the states of affairs, particularly actions, were to be arbitrarily excluded from the domain of situated consequential evaluation.

Sometimes the demand for excluding particular components of the state of affairs comes from the claim that they are somehow not really part of the state of affairs *at all*. For example, it has been claimed that an action performed is not a part of a state. The fact that to do something is to make something happen is an elementary enough point, and yet not only has this been denied in many anti-consequentialist arguments, that denial has ended up having pivotal importance in the strategy of such arguments. In his earlier book, *The*

---

[17] "Utilitarianism and Welfarism," "Rights and Agency," "Well-being, Agency and Freedom: The Dewey Lectures 1984," and *On Ethics and Economics* (Malden, MA: Blackwell, 1987).

Rejection of Consequentialism, Scheffler[18] not only develops this line of reasoning, but also quotes a "personal communication" from Thomas Scanlon allegedly making this point. Scanlon is quoted as saying: "the distinctive moral badness of killings compared with other deaths exists, if at all, only in an agent-centered morality focusing on what you *do*. Once we move to the moral appraisal of what *happens*, the differential treatment of killings and otherwise equally undesirable deaths becomes much less plausible" (*ibid.*, p. 109). A state of affairs certainly does consist of what has happened, but if an action *x* has been performed by person *i*, then it certainly has happened that *i* has done *x*. It would require an odd surgery to "remove" this occurrence from the description of the state of affairs, and that would be an epistemically impoverishing and morally debilitating exclusion.

If Arjuna kills someone, it would be peculiar to deny that such a killing has occurred. And in doing a moral evaluation of what has happened, Arjuna can hardly overlook the fact that this killing has happened, and that he himself has played the active part in this event. It is a part of history, and it is an event that Arjuna in particular must take seriously. The discipline of consequential evaluation of states of affairs does not offer a special exemption to anyone to overlook what she has done on the ground that we must be concerned only with what has *happened*, not what has been *done*.

There are reasons to pause to consider this issue further, since Scanlon has developed this argument more fully in his later work. One crucial issue, among others, is Scanlon's argument that "the negative intrinsic value attached to morally undesirable actions is impartial—that is to say that an action that has this disvalue is something that everyone has reason to prevent" (*op. cit.*, p. 83). The latter part of the claim (not the former) need not be denied even within broad consequentialism (of the kind I have been exploring), and it can indeed be argued that everyone has a reason to prevent, say, a murder from occurring. The real issue is not that, but whether everyone has the *same* reason, with the *same* force—the murderer no more than anyone else—to avert that murder. The idea that the state of affairs must be evaluatively viewed in the same way by everyone sharing the same ethics, despite their different roles (in, say, having murdered someone *or* not) is arbitrary as well as implausible.

Scanlon points out that "if we are to make sense of the idea of a deontological prohibition against killing by assigning intrinsic disvalue to acts of killing, this disvalue must not be impartial but what is sometimes called agent-relative disvalue" (*ibid.*). With this proposition I am basically in agreement, except that I would go beyond his claim and argue that this disvalue must be agent relative not just in the case of deontological *prohibitions,* but also in other—less extreme—cases of giving shape to deontological concerns with actions and agencies. Indeed, the scope of agent-relative judgments need not be confined only to "prohibitions," which is a specially limiting way in which some deontological systems accommodate agent-relative values. In broad deontology as well as in broad consequentialism, there may be agent-relative variations that do not take the form of a prohibition. For example, in Bernard Williams's illuminating analysis of the now-famous case as to whether it would be right for Jim to follow the utilitarian logic to kill one captive if it would save the lives of the other nineteen, Williams's main concern is with identifying and explaining the important considerations that utilitarianism leaves out. His conclusion is not so much that the utilitarian would come to the wrong conclusion in this case, but, as Williams puts it, "if (as I suppose) the utilitarian is probably right in this case, that is not to be found out just by asking the utilitarian's questions."[19] Deontological prohibition is a very special case of agent-relative judgments.

---

[18] New York: Oxford, 1982.

In a broad consequential system, the agent-relative variation of valuation may distinguish the special badness of killing vis-à-vis allowing a death to occur, and it is only if the extra disvalue of killing is infinitely negative for the murderer, then (and only then) the agent-relative consequential system would impose a total prohibition on killing. The distinction is of importance in this analysis, and it has some relevance even to Scanlon's further argument that "agent-relative reason" cannot "flow from the disvalue of an action, understood as an event" (*op. cit.*, p. 83), a proposition that neatly states our main point of disagreement. I have already discussed why responsibility for the consequences of one's choice cannot be suspended to provide a special exemption to the chooser for the event that takes the form of a murder that he commits. But among our differences is the emphasis that Scanlon seems to put on the principle of "deontological prohibitions."

Scanlon rightly points out that someone "who accepts this principle [of deontological prohibitions] therefore does not need to appeal to the 'negative intrinsic value' of killing in order to explain why she does not do what is necessary to save the greater number" (*op. cit.*, p. 84). This is indeed so, but it applies only to the extreme case of total prohibition (and does not come to terms with, for example, Williams's rich analysis of Jim's dilemma, nor with his balanced overall conclusion). This is adequate enough if one must never kill—not even to prevent a genocide of a million people. To allow other—less extreme—possibilities, deontological theory would have to take a broader form than constraint-based deontology permits. And then the approach through the special badness (what is unprepossessingly described as "negative intrinsic value") of killing is not such a patently redundant idea.

---

[19] "A Critique of Utilitarianism," p. 117. Williams's critique is primarily against utilitarianism, but he relates this to his rejection of consequentialism in general, without characterizing his preferred alternative in the form of deontological prohibitions.

I shall not go further into this contrast here.[20] But I must also point out that there are other parts of the state of affairs (other than the actions in general and the agent's own actions in particular) which are sometimes left out of the accounting of the state of affairs in making the claim that they lie beyond the reach of consequential evaluation. This is often claimed about the *motivation* of a person in undertaking an action, in which standard utilitarian ethics has no direct interest. It too can be important, however, in broader consequential systems.

Indeed, motivation can be seen to be part and parcel of what is happening and may well be taken into account when a state of affairs is judged. Consider, for example, a person who says: "You say I should be satisfied with the fact that he has apologized; I agree that he has used words of apology, but what has not actually happened is his really feeling sorry and apologizing genuinely; he has just apologized to escape chastisement, and this is not good enough." The distinction being made (about what has "actually happened") concerns the motivation underlying an action.

A state of affairs is informationally rich. There is no particular reason to insist on an impoverished account of a state of affairs in evaluating it. Also, the reach of consequential reasoning can incorporate *processes* of choice, and not merely the narrowly defined ultimate outcomes. In the context of decision theory and rational choice, I have argued for the importance of paying particular attention to "comprehensive outcomes" (including actions undertaken, processes involved, and the like *along with* the final outcomes), instead of confining attention to only the "culmination outcome" (what happens at the very end).[21] The distinction can be very central to certain problems in economics, politics, sociology, and in the general theory of rational decisions and games. As it happens, the distinction is also crucial in assessing the reach of consequential reasoning. If, for example, a presidential candidate were to argue that what is really important is not just to win the forthcoming election, but "to win the election fairly," then the outcome recommended is a *comprehensive outcome*, which includes a process consideration (not just the culmination outcome of winning the election—no matter how).[22] The appraisal of comprehensive outcomes can be an integral part of the assessment of states of affairs, and thus a crucial building block in consequential evaluation.

---

[20] See also my "Rights and Agency," "Evaluator Relativity and Consequential Evaluation," "Well-being, Agency and Freedom: The Dewey Lectures 1984," and *On Ethics and Economics.*

VI. RIGHTS AND OBLIGATIONS

The type of consequentialist evaluation that is being explored here can adequately withstand various standard critiques of consequentialism which can be found in the literature. I shall come presently to that defensive issue. But before that, I want to look at a more positive aspect of the outlined consequentialist approach, in particular its extensive reach. I shall illustrate this reach by showing how the proposed consequential framework may be able to accommodate the concept of moral (rather than legal) rights in general, and "human rights" in particular (a concept that is frequently invoked in practical discussions in the contemporary world).

There is, of course, no necessity to invoke the idea of rights in any form in making use of the broad consequential approach I have been exploring here, but it is also the case (as it were, a "further fact") that rights, including human rights, can be placed within this framework. One way of thinking about this issue is to separate out two aspects of a standard concept of rights: (1) it aims at the *freedom* of the right-holder to do certain things or achieve some conditions, and (2) it demands some correlate *obligations* on the part of others (which can take the form of noninterference *or* of positive assistance) to help in the realization of this freedom by the right-holder. For example, person *A*'s right not to be assaulted concerns both (1) the freedom of

---

[21] The nature and the importance of the distinction between "culmination outcomes" and "comprehensive outcomes" are more fully analyzed in my "Maximization and the Act of Choice."

[22] Even if it were the case that the presidential candidate actually wants—no matter what he says—to win the election by fair means or foul, the argument presented by him is for the excellence of "winning fairly." The point is not about the actual preference of the person (even less about the likelihood of such a preference), but only about a valuational judgment on the importance of a comprehensive outcome of this kind.

A to avoid being assaulted, and (2) the obligation of others not to assault *A*. (It may also include a general obligation of others to try to prevent a third person from assaulting *A*, but that involves some complication, which I shall address later.) A comprehensive outcome can be judged to have been worsened on *each* of these—distinct but interrelated—grounds if, say, person *B* were to assault *A*: this would be a nonfulfillment of *A*'s right and a nonperformance of *B*'s duty. The right and the obligation are not independent events, but they may not be invariably locked to each other (for example, *B*'s serious attempt at assaulting *A* may fail, and yet the badness of *B*'s having tried to assault would still be a factor to consider in assessing the comprehensive outcome). The realization (or not) of substantive freedoms as well as taking adequate account (or not) of obligations can *both* figure in ethical accounting within this broad consequential framework.

It may be worth scrutinizing how exactly does this incorporation of the fulfillment of rights and of correlated duties *within* a consequential framework differ, on the one hand, from *rights-independent consequential evaluation*, for example, in utilitarianism, and, on the other, from *rights-inclusive nonconsequential procedures*, for example, in constraint-based libertarianism.[23] Utilitarians do not include the realization of freedoms, or the fulfillment of rights or duties, among the valued objects at all, and so there is a fundamental gap here. Of course, rights or duties may be instrumentally valued by utilitarians for what they can do to promote utilities, but their fulfillment or violation does not, by itself, make the states better or worse in utilitarian accounting.

Turning now to libertarianism, it does of course pay much direct attention to rights and duties, but only as constraints that restrict what others are permitted to do. No one should assault *A*, and if *B* does carry out an assault on *A*, then he has violated a constraint that he should have obeyed. But in pure libertarian theory, it is not claimed that the state of affairs has turned worse as a result of that assault. The badness of the assault is not the issue in that approach, only its wrongness is.[24]

So neither the utilitarian nor the libertarian would argue—or insist (as a rights-inclusive consequentialist would)—that the assault was, in itself, a bad thing to happen. In judging the state of affairs, utilitarianism would turn only to utilities.[25] The libertarian is not concerned

---

[23] See Nozick.

[24] On this, see Nozick, chapter 7.

[25] The assault, not seen as in itself bad, gets utilitarian attention only if—and only to the extent that—utilities are reduced overall as a result of the assault (in fact, of

(at least in this exercise) with judging the state of affairs at all. In contrast with both these approaches, consequential evaluation that takes note of freedoms, rights, and obligations—and their violation—would argue that bad things have happened precisely because someone's freedom has been breached, and some rights and duties have been violated.[26]

VII. HUMAN RIGHTS AND IMPERFECT OBLIGATIONS

I turn now to human rights in particular. These are rights that individuals are supposed to have not by virtue of their citizenship of any particular country, but because of their status as human beings. These may or may not be legislated rights, but in so far as they are valued, that valuation can include the importance that is attached to the relevant persons' freedoms and also the responsibility that others have—irrespective of citizenship, nationality, and other denominations—to help this person to attain these freedoms. If others can help, then there is a responsibility that goes with it. Even if it is not specified who will have to do what to help the victimized person, there is a general need for a responsible consequence-evaluating agent to consider her general duty to help others (when reasonably feasible).[27]

Some of these obligations tend to be more fully specified than others. For example, the insistence that the state must not arbitrarily arrest a person (if that is seen as a human right, no matter whether the laws of the land prohibit arbitrary arrest or not) is more exactly characterized than, say, the demand that a person has a right to help in escaping starvation (it demands, generally though imprecisely, that all those who are in a position to help should try to consider what they can do to prevent this eventuality from occurring). Demands of human rights can take either form, *with or without* pinpointing specific duties that the obligation bearers are asked to accept.

In fact, even the fulfillment or violation of precisely specified obligations can go with imperfect obligations of others to help in a general way. Since gross violations of serious obligations can be seen as making the states of affairs worse, even others not directly involved in the violation of a perfectly specified obligation (for example, person *A*'s being assaulted by person *B*) may have a general duty to help (in this case, to try to prevent *B*'s assault on *A*). This duty, through a consequential link, may be rather inexactly specified (tell-

---

course, the overall utility picture can, in many circumstances, move in precisely the opposite direction). On this, see my "Rights and Agency."

[26] On this, see my "Rights and Agency."

[27] This requirement and its extensive practical implications are discussed in my *Development as Freedom* (New York: Knopf; and New York: Oxford, 1999).

ing us neither who must particularly take the initiative, nor how far he should go in doing this general duty), but this loosely formulated obligation–Immanuel Kant[28] would call it an "imperfect obligation"—to help may nevertheless be seriously considered by (and be influential with) responsible people. Indeed, neglect or disregard of such general obligations may be plausibly seen as a bad thing to happen, and can be taken into account in evaluating states of affairs. For example, if a person were severely assaulted in full view of others and her cries for help were completely ignored, it is not unreasonable (in discussing what was so bad about that event) to argue that three bad things happened: (1) the victim's freedom was violated, and so was her right not to be assaulted; (2) the assaulter transgressed the immunity that others should have from intrusion (in this case, a violent intrusion) and violated his duty not to assault others; and (3) the others who did nothing to help the victim also transgressed their general—and imperfect—obligation to help others (which they could reasonably be expected to provide). They are interrelated failings, but distinct from each other.

It is important to see that in linking human rights to both perfect and imperfect obligations, there is no suggestion that the right-duty correspondence be denied. Indeed, the binary relation between rights and obligations can be quite important, and it is precisely this binary relation which separates out human rights from the general valuing of freedom (without a correlated obligation of others to help bring about a greater realization of human freedom). The question that remains is whether it is adequate for this binary relation to allow *imperfect* obligations to correspond to human rights, without demanding an exact specification of who will have to do what, as in the case of legal rights and specified perfect obligations.

This question is indeed the basic source of some of the skepticism that greets assertions of human rights without exact identification of particular perfect obligations that would lead to the realization of these rights. In the absence of such perfect obligations, demands for human rights are often seen just as loose talk (perhaps wonderful loose talk, but loose talk nevertheless). One question that motivates some of this skepticism is: How can we be sure that rights are, in fact, realizable unless they are matched by full specification of who will do what? Indeed, some do not see any sense in a right unless it is balanced by what Kant called a "perfect obligation"—a specific duty of a particular agent for the actual realization of that right. Onora

---

[28] *Critique of Practical Reason*, L.W. Beck, trans. (New York: Bobbs-Merrill, 1956).

O'Neill has argued powerfully in this direction in her book, *Towards Justice and Virtue*[29]:

> Unfortunately much writing and rhetoric on rights heedlessly proclaims universal rights to goods and services, and in particular "welfare rights," as well as to other social, economic and cultural rights that are prominent in international Charters and Declarations, without showing what connects each presumed right-holder to some specific obligation-bearer(s), which leaves the content of these supposed rights wholly obscure. This obscurity has been a scene and source of vast political and theoretical wrangling. Some advocates of universal economic, social and cultural rights go no further than to emphasize that they *can* be institutionalized, which is true. But the point of difference is that they *must* be institutionalized: if they are not there is no right (*ibid.*, pp. 131-32).

O'Neill goes on to identify the constructive role of "virtue," instead of rights, since rights, in this view, make no sense when unaccompanied by specified perfect obligations. For example, so long as the putative right to food is not matched up with some particular person's or agent's specific obligation, the hungry do not really, in this reckoning, have a right to food at all. On the other hand, it would still be virtuous to feed the hungry.

This line of reasoning has considerable persuasive power, and O'Neill has done much to explore and illuminate the territory of virtue.[30] We can, however, ask: Why insist on the absolute necessity of co-specified perfect obligation for a putative right to qualify as a real right? Certainly, a perfect obligation would help a great deal toward the realization of rights, but why cannot there be *unrealized* rights, even rights that are hard to realize? We do not, in any obvious sense, contradict ourselves by saying: "These people had all these rights, but alas they were not realized." Something else has to be invoked to jump from pessimism about the *fulfillment* of rights, all the way to the *denial* of the rights themselves.

This distinction may appear to be partly a matter of language, and it might be thought that O'Neill's rejection could be based on how the term 'rights' functions in common discourse. Perhaps O'Neill is pointing to language use and its discipline? But in fact the term 'rights' is used much more widely than O'Neill demands. Indeed, it is a significant part of her claim that the term 'rights' is often used quite wrongly; she is not relying on language usage—rather going *against* it. The language of rights, in public as well as private discussion, often proceeds without any clear identification of perfect obligations. Rather, the claim to which O'Neill gives expression must be based on the presumption that using rights in this way—what is seen by her as a wrong way—is full of internal problems which can be remedied only by more exacting specification of rights with correlated obligations. O'Neill's point, then, is not usage but the cogency of usage.

---

[29] New York: Cambridge, 1996.

[30] Indeed, our understanding of the role of virtue is very significantly enriched by O'Neill's constructive analysis of its extensive reach presented in *Towards Justice and Virtue*, but also by the far-reaching investigation in her earlier book—*Faces of Hunger: An Essay on Poverty, Justice and Development* (London: Allen and Unwin, 1986)—of the role of sympathetic ethics and humane politics, in the remedying of hunger in the contemporary world. It is also important to note that O'Neill does not oppose "welfare rights," but wants them to be institutionalized (otherwise, "there is no right").

I would, however, resist the claim that any use of rights except with co-linked perfect obligations must lack cogency. Indeed, it can be claimed that the problem that O'Neill perceives arises partly from her implicit attempt to see the use of rights in political or moral discourse through a close analogy with rights in a legal system, with its demand for specification of correlated duties. In contrast, in normative discussions rights are often championed as entitlements or powers or immunities that it would be good for people to have. Human rights are seen as rights shared by all—irrespective of citizenship—and the benefits of which everyone *should* have. The claims are addressed generally—in Kant's language "imperfectly"—to anyone who can help. Even though no particular person or agency has been charged with bringing about the fulfillment of the rights involved, they can still be very influential. Even if it is not feasible that everyone can have the fulfillment of their rights in this sense (if, for example, it is not yet possible to eliminate undernourishment altogether), credit can still be taken for the *extent* to which these alleged rights are fulfilled. The recognition of such claims as rights may not only be an ethically important statement, it can also help to focus attention on these matters, making their fulfillment that much more likely—or quicker.[31]

This is indeed the form in which many major champions of rights-based thinking have tried to use the idea of rights. The plausibility and advantage of thinking in terms of rights was brought out early enough by Tom Paine and Mary Wollstonecraft, who were contemporaries (indeed, Paine's *Rights of Man* and Wollstonecraft's *The Vindication of the Rights of Women* were published in the same year,

---

[31] These issues are discussed in my *Development as Freedom*, chapter 10, which also critically examines some specific arguments against the idea of human rights (related respectively to their "legitimacy," "coherence," and "cultural basis").

1792).[32] To dismiss this powerful literature by claiming that they were making some kind of an elementary mistake would seem to be based on the odd assumption that they were talking about something other than what they were talking about. Or, alternatively, that they *should be* talking about something other than what they wanted to discuss.

In fact, it is not in general cogent even to presume that, if an inexactly specified human right is important, then it must be excellent to try to legislate it into a precisely specified right and to "institutionalize" it as an exact right. For example, a wife's "human right" to be consulted in family decisions even in a traditionalist society (where these decisions are typically taken by the husband on his own) may be important to emphasize and even to fight for. And yet a great defender of this human right (and its far-reaching ethical and political implications) might agree that it is not sensible to make this human right into a legal right (perhaps with the husband's being arrested and locked up if he were to fail to consult his wife). The importance of human rights need not lie only in their being putative proposals for legislation and institutionalization. Human rights can have their own domain of influence and importance *without* being parasitic on—or even being wisely reinforced by—subsequent legislation.

The issue of cogency cannot be decided by obliterating the difference between two dissimilar types of usage of the idea of rights, and by insisting that cogency in one usage must mirror cogency in the other. If the fulfillment of rights is a good thing to happen—the more the better—as it would be seen in a consequential perspective, the condition of full feasibility cannot be a condition of coherence. An analogy with utilitarianism may be helpful. Utilitarian reasoning admits the possibility and excellence of everyone enjoying what Frank Ramsey[33] called "bliss." This is among the alternatives that are considered—indeed better than any other—but it is not, of course, claimed that the utilitarian exercise would lack cogency if bliss for all is not realized, nor be realizable (indeed quite the contrary). Indeed, Ramsey did not presume that "bliss" would be realizable, but this did not reduce the cogency of his exercise of minimizing the aggregate *shortfall* from bliss. Similarly, rights-inclusive objectives in a system of consequential evaluation can accommodate certain rights the fulfillment of which would be excellent but not guaranteed, and we can still try to minimize the shortfall.

---

[32] Paine (London: Symonds, 1792); and Wollstonecraft (London: Johnson, 1792).

[33] *Foundations of Mathematics and Other Logical Essays* (London: Kegan Paul, 1931).

VIII. CONSEQUENTIAL EVALUATION AND INTERDEPENDENCE

One of the positive merits of consequential evaluation is the discipline that it imposes and the decisional issues it makes each agent face. If rights are embedded in a system of consequential evaluation, we are immediately led to questions about the badness of the violation of rights, or the goodness of their fulfillment, and it has implications on good choices and good actions. These consequential implications suggest how different rights may be assessed vis-à-vis each other, and how the priorities between them may be systematically appraised within a coherent consequentialist framework. Those who take rights as side constraints tend to argue against any such balancing, and seem to be dead set against endorsing the infraction of any right—no matter how minor—for the sake of preventing a more important violation of another—more crucial—right.[34]

Indeed, we often have to face issues that explicitly or implicitly involve the balancing of different rights, or the balancing of rights against other good things. Suppose you can prevent a murder from taking place or a rape from occurring by appearing on the scene. But to do that you need to use a motor car and you do not have one. Someone you know does have a car just there and you can find your way into it, but you know that he would be most unwilling to let you use his car for this purpose. Do you have a reason to violate his right over his own car to prevent the violation of life and liberty which would occur if the murder or rape were to take place (if you would fail to show up)?

It is possible that not everyone will agree as to what should be done, but it is not absurd to think that you have a reasonable case for violating the minor right of the car owner to prevent a bigger violation of more important rights of the person who is threatened with rape or murder.[35] But this will be directly against obeying the side constraint of not violating anyone's rights, including the car owner's, and in that system of constraint-based deontology, you must not be moved by the relative gravity of different violations of rights, or by other consequences. On the other hand, with a consequential approach, such comparisons can be part and parcel of taking responsibility for the consequences of one's actions. Indeed, in that approach, refusing to consider the relative importance of the different violations of rights can well count as a significant abdication of responsibility.

---

[34] Nozick (*op. cit.*) argues against such trade-offs between rights, which he sees as "utilitarianism of rights." Utilitarianism of rights would indeed be contradictory, but consequentialism of rights need not be. Notice that there are three distinct though interrelated departures from the libertarian framework: (1) violation of rights are included among the consequences; (2) there can be "trade-offs" between rights; and (3) there can be "trade-offs" between the goodness of rights fulfillment and other good consequences.

[35] I have discussed this issue in "Rights and Agency."

The need to examine the relative importance of different consequences (including well-beings, freedoms, rights, and so on) that may compete with each other in evaluative assessment does arise in many different contexts. We live in an interdependent world in which the realization of our respective freedoms interconnects in a variety of ways, and we cannot treat them each as an isolated island. The discipline of consequential evaluation forces us to take responsibility for our choices, since our actions influence other people's freedoms and lives as well as our own. The reach of our responsibility includes asking certain questions, such as those concerning the relative importance of different rights or freedoms the realization of which may impinge on each other.

But situations of interdependence have also been used to present an important class of arguments that resist consequential reasoning, and indeed the merits of deontological reasoning is often illustrated by invoking interdependences of a kind in which—it has been argued—consequential evaluation would lead to just the wrong decision. Commenting on an earlier paper of mine, "Rights and Agency," Philippa Foot[36] notes that embedding the fulfillment or violation of rights within a consequential structure "would help to solve some problems," but leave others untouched:

> ...this "goal rights" system fails to deal with certain other examples of actions that most of us would want to call wrong. Suppose, for instance, that some evil person threatens to kill or torture a number of victims unless we kill or torture one, and suppose that we have every reason to believe that he will do as he says. Then in terms of their total outcomes (again consisting of the states of affairs made up an action and its consequences) we have the choice between more killings and torturings and less, and a consequentialist will have to say that we are justified in killing or torturing the one person, and indeed we are morally obliged to do it, always supposing that no indirect consequences have tipped the balance of good and evil. There will in fact be nothing that it will not be right to do to a perfectly innocent individual if that is the only way of preventing another agent from doing more things of the same kind (*ibid.*, p. 226).

---

[36] "Utilitarianism and the Virtues," *Mind*, xciv, 374 (1985): 196-209; reprinted in Scheffler, *Consequentialism and Its Critics*.

This kind of example has also been considered by Williams, Scheffler, Thomas Nagel, Derek Parfit,[37] among others, to present a similar rejection of consequentialism.

Does this argument against the plausibility of consequential evaluation work? It would indeed work well against a *combined system* that joins consequentialism with further requirements of *insensitivity*, in particular the wholly additional—and (I have argued) quite unjustified—demand that evaluation of consequences be insensitive to the nature of the act and the position of the agent (this further restriction is imposed, as was discussed earlier, by "welfarism" in the utilitarian system). But it offers nothing against consequential evaluation with situated valuations, for reasons that were also discussed earlier. When I am evaluating a choice for me, with one option being my killing an innocent person, I can hardly see it in the same way as another killing that may be occurring somewhere else without my being involved in it. No matter what we ultimately decide in very complex cases in which my killing one person may save many lives, perhaps even a million lives (there are indeed arguments on both sides), the two happenings, namely, (1) my killing someone and (2) a death unrelated to me occurring, are most emphatically not the same. They are descriptively quite different, and that difference should be of considerable normative significance when I come to evaluate the consequences of my choices. Why must we assume that they should be "taken to be the same," when they clearly are not? Indeed, it is precisely the badness of killing innocent people that added force to Arjuna's *consequential* argument against Krishna's defense of the *deontological obligation* to fight for a just cause.

IX. A CONCLUDING REMARK

The starting point of this article is the importance of taking responsibility for one's choices and their consequences. This provides the motivation behind the discipline of consequential evaluation as outlined here. In that light, I have scrutinized the relevance and reach of consequential evaluation without amalgamating it with extraneous demands such as welfarism (imposed by utilitarianism), which insists that all consequences other than utility consequences be ultimately ignored. The responsibility of choice applies to the choice of evaluative perspective as well as that of actions and conduct. I have tried to analyze both (1) the general demands of consequential evaluation, and (2) more specifically, features of a particular class of consequential evaluations that are distinctly nonutilitarian.

---

[37] Williams, "A Critique of Utilitarianism," in Smart and Williams; Nagel, "The Limits of Objectivity," in S. McMurrin, ed., *Tanner Lectures on Human Values*, Volume I (New York: Cambridge, 1980); Scheffler, *The Rejection of Consequentialism*; Parfit, *op. cit.*

Some of the *alleged* limitations of consequential reasoning can be seen to be generated not by the discipline of consequential evaluation itself, but by additional assumptions—entirely separate and by no means necessary—with which a consequential approach is frequently combined. Three supplementary assumptions, extraneous to the general approach of consequential evaluation, are especially limiting: (1) the insistence on a complete ordering of all states of affairs (thereby confounding *maximization* with *optimization*); (2) the removal of actions, motives, processes, and other features of states of affairs from artificially denuded concepts of states of affairs (thereby ignoring the breadth of *comprehensive* outcomes in favor of the narrowness of *culmination* outcomes); and (3) disregarding the positionality of the agents vis-à-vis the actions and the resulting states of affairs (thereby replacing *situated valuations* by selectively blind assessment). There is, in fact, no necessity to combine consequential evaluation with these supplementary assumptions, and then to damn consequential evaluation for the company it is thus forced to keep.

Within a broad consequential framework, ethical or political rights can be seen to be less narrowly specified than the formalities of legal rights. The alleged need to specify "perfect obligations" (specifying exactly who will do what) to correspond to rights, which are sometimes taken to be essential for making any sense of rights, are overly restrictive. In contrast, looser forms of obligations (sometimes called "imperfect obligations"), which point to the need for others to pay serious attention to their responsibilities in preventing harm and the violation of freedoms and rights, can have an extensive role in making us understand such concepts as "human rights" and "basic civil rights." They need not be dismissed as just loose talk.

As it happens, communication and debating on rights and duties are full of serious difficulties and natural hurdles, because of substantive differences in our values and priorities. We need not add to these real problems a spurious one arising from a refusal to see the way rights and duties may be cogently understood in consequence-oriented evaluation. The political reasoning underlying the idea of human rights calls neither for misconceived dismissal, nor for good-humored tolerance of well-meaning but allegedly confused activists, but for a better understanding of the conceptual basis of these claims. Consequential evaluation has something to offer here, as it has in many other fields as well.

AMARTYA SEN

Trinity College, Cambridge

TWO DOGMAS OF BELIEF REVISION*

W. V. Quine's epoch-making paper "Two Dogmas of Empiricism"[1] sealed the fate of logical positivism by undermining two presumptions of classical as well as modern empiricism: the idea that there is a sharp distinction between analytical judgments (knowledge of meanings) and synthetical judgments (knowledge of facts), and the idea that every meaningful sentence can be reduced to a construction upon observation reports. The most interesting feature for my purposes here, however, is that Quine closed his paper with a beautiful section on "empiricism without the dogmas" which deals almost exclusively with a topic that would nowadays be called *theory change*, or *belief revision*. He paints a picture of how we should, and for the most part do, accommodate the scientific lore if we meet with recalcitrant experiences. According to Quine's pragmatist approach, theory revision is a matter of choice, and the choices are to be made in such a way that (a) the resulting theory squares with the experience, (b) it is simple, and (c) the choices disturb the original theory as little as possible.

Quine's picture is mainly metaphorical, and he never intended to turn it into a formal theory of theory change. The situation today is different. At least since the seminal work of Carlos Alchourrón, Peter Gärdenfors, and David Makinson[2] (*AGM*) in the early 1980s, there has been a clear-cut paradigm of how logically to formalize theory change. The original AGM approach has turned out to be limited in various ways, and many extensions and revisions of it have been developed over the past two decades. I do not want to find fault with any of the formal theories of belief revision which are currently being advocated. My aim

---

* A preliminary version of this paper was presented at the Third Conference on Logic and the Foundations of the Theory of Games and Decisions (Torino, December 1998). I am grateful to Eduardo Fermé, Oliver Schulte, and in particular to Isaac Levi and David Makinson for a number of very valuable comments on earlier versions of this paper.

[1] *Philosophical Review*, LX (1951): 20-43; reprinted in Quine, *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 20-46.

[2] See Alchourrón, Gärdenfors, and Makinson, "On the Logic of Theory Change: Partial Meet Contraction Functions and Their Associated Revision Functions," *Journal of Symbolic Logic*, L (1985): 510-30; Gärdenfors, *Knowledge in Flux: Modeling the Dynamics of Epistemic States* (Cambridge: MIT, 1988); Gärdenfors and Rott, "Belief Revision," in D.M. Gabbay, C.J. Hogger, and J.A. Robinson, eds., *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume IV: Epistemic and Temporal Reasoning* (New York: Oxford, 1995), pp. 35-132; and Sven Ove Hansson, *A Textbook of Belief Dynamics: Theory Change and Database Updating* (Boston: Kluwer, 1999).

0022-362X/00/9709/503–22

here is rather to call into question what has been taken to be *the* principal idea behind the current theories of belief revision: the idea of *informational economy.* This idea is basically the same as Quine's criterion (c), often called the *principle of minimum mutilation,* or also *(doxastic* or *epistemic) conservatism.* I shall try to cast doubt not on the principle itself, but on its being *the* philosophical background that makes AGM-style theories of belief revision intelligible in the first place.

The principle of informational economy tells us that we should not give up our beliefs beyond necessity. It is not spelled out clearly here what 'necessity' means. It might be suggested that we can satisfy the principle of minimal change in an ideal way simply by not changing our beliefs at all.[3] But that is certainly not intended. There are basic requirements for belief revision which have to be satisfied, and only against the background of these conditions can we apply the minimal-change principle. One of the basic requirements that will be left unquestioned here is that revisions should be *successful* in accommodating new information: the kind of revision that we consider is such that incoming input always gets accepted. That this is too strict as a general constraint for a truly realistic theory of belief change has been recognized by many writers and is taken account of in numerous recent papers on "non-prioritized belief revision."[4]

The second basic requirement I am going to impose is that agents should be taken to be ideally competent regarding matters of logic. They should accept all the consequences of the beliefs they hold (that is, their set of beliefs should be *logically closed*), and they should rigorously see to it that their beliefs are *consistent* (it is just the task of belief-revision theory to give an account of how they can manage that their beliefs remain consistent). As stressed by researchers in belief *base* revision and paraconsistent logics, respectively, the insistence on closure and consistency may be regarded as unrealistic in general. The case is not as clear as with the success condition, however, since there are at least two interpretations of belief which make the logical restrictions compelling: beliefs as attitudes ascribed by a third person, and beliefs as commitments.[5]

---

[3] This means pursuing Charles S. Peirce's "method of tenacity"; see his "The Fixation of Belief," in H. S. Thayer, ed., *Pragmatism: The Classic Writings* (New York: New American Library, 1970), pp. 61-78, here pp. 68-70. Thanks to Isaac Levi for raising this point.

[4] See, in particular, Hansson, ed., Special Issue on Non-Prioritized Belief Revision, *Theoria,* LXIII (1999): 1-134.

[5] In AGM theory, the success condition has precedence over the consistency condition, since a revision by a contradiction leads to an inconsistent belief set. But this seems to be a contingent rule for a logical limiting case rather than a matter of

I shall assume from now on that these *basic requirements*—that belief sets should be closed and that revisions should be successful and lead to consistent belief sets—form the background against which questions of informational economy or minimal change are discussed. Agents who do not change their beliefs at all can perhaps be called very "economical" in the administration of what they currently possess, but they are not able to interact successfully with the world. "Do nothing!" is not a feasible option for our cognitive life.

I

Informational economy has been proclaimed to lie at the basis of belief change from the very beginning of the systematic study of belief revision. The idea comes in two distinct versions:

(1) When accepting a new piece of information, an agent should aim at a minimal change of his old beliefs.
(2) If there are different ways to effect a belief change, the agent should give up those beliefs which are least entrenched.

These maxims have been accompanying belief-change theory since its inception, and they were repeated in the literature time and again. Here are some relevant references.

> Ad (1).
>
> The concept of contraction leads us to the concept of *minimal change of belief*, or briefly, *revision* (Makinson[6]).
>
> The criterion of informational economy demands that as few beliefs as possible be given up so that the change is in some sense a *minimal* change of *K* to accommodate for *A* (Gärdenfors, *op. cit.*, p. 53).
>
> The amount of information lost in a belief change should be kept minimal (Gärdenfors and Rott, *op. cit.*, p. 38).
>
> At the center of the AGM theory [of theory change] are a number of approaches to giving formal substance to the maxim [of minimal mutilation: keep incisions into theories as small as possible!] (André Fuhrmann[7]).
>
> The hallmark of the AGM postulates is the principle of minimal belief change, that is, the need to preserve as much of earlier beliefs as possible and to add only those beliefs which are absolutely compelled by the revision specified (Adnan Darwiche and Judea Pearl[8]).

---

considered philosophical decision. I assume here that new information is always consistent in itself.

[6] "How to Give It Up: A Survey of Some Formal Aspects of the Logic of Theory Change," *Synthese*, LXII (1985): 347-63, here p. 352.

[7] *An Essay on Contraction* (Stanford: CSLI, 1997), p. 17.

[8] "On the Logic of Iterated Belief Revision," *Artificial Intelligence*, LXXIX (1997): 1-29, here p. 2.

> Ad (2).
>
> When a belief set K is contracted (or revised), the sentences in K that are given up are those with the *lowest* epistemic entrenchment (Gärdenfors, *op. cit.*, p. 87).
>
> The guiding idea for the construction is that when a knowledge system K is revised or contracted, the sentences in K which are given up are those having the *lowest* degrees of epistemic entrenchment (Gärdenfors and Makinson[9]).
>
> In so far as some beliefs are considered more important or entrenched than others, one should retract the least important ones (Gärdenfors and Rott, *op. cit.*, p. 38).[10]
>
> ...when it comes to choosing between candidates for removal, the least entrenched ones ought to be given up (Fuhrmann, *op. cit.*, p. 24).
>
> A hallmark of the AGM theory is its commitment to the principle of *informational economy*: beliefs are only given up when there are no less entrenched candidates.... If one of two beliefs must be retracted in order to accommodate some new fact, the less entrenched belief will be relinquished, while the more entrenched persists (Craig Boutilier[11]).

Although (1) and (2) look quite different and it is not immediately clear how they fit together, there is a result that appears to show that they are even "at root identical"[12] and that they can therefore be viewed as two incarnations of a unified idea of informational economy. The result mentioned maps onto one another two prominent types of belief-change constructions investigated by Alchourrón, Gärdenfors, and Makinson: belief changes obtained by partial meets of maximal nonimplying sets of beliefs,[13] and belief changes based on relations of epistemic entrenchment.[14] These two methods are generally taken to be closely associated with (1) and (2), respectively, and so it is both surprising and pleasing to find that they can be proved equivalent in a rather strict sense by directly relating the underlying relations used.[15]

---

[9] "Revisions of Knowledge Systems Using Epistemic Entrenchment," in Moshe Vardi, ed., *TARK '88: Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge* (Los Altos: Kaufmann, 1988), pp. 83-95, here p. 88.

[10] Similarly, Gärdenfors, "Belief Revision: An Introduction," in his (ed.) *Belief Revision* (New York: Cambridge, 1992), pp. 1-28.

[11] "Iterated Revision and Minimal Change of Conditional Beliefs," *Journal of Philosophical Logic*, xxv (1996): 263-305, here pp. 264-65.

[12] As, by the way, Quine (section 5) claimed the two dogmas of empiricism were.

[13] Introduced by Alchourrón, Gärdenfors, and Makinson.

[14] Introduced by Gärdenfors and Makinson.

[15] The relevant result is presented in Section 4 of my "Two Methods of Constructing Contractions and Revisions of Knowledge Systems," *Journal of Philosophical Logic*, xx (1991): 149-73. I shall return to this topic in subsection III.B and argue that the interpretation just given is incorrect. Criteria (1) and (2) may even be pulling in opposite directions; compare footnote 32.

The overall picture that we have now got of the AGM approach seems nice and harmonious. But I want to argue that *maxims (1) and (2) are a travesty of the principles that have actually been followed in the traditional theories of belief revision.* The philosophical underpinnings of the prevailing theories of belief change are not at all what the folklore would like to have them.

As will become clear later on, I call (1) and (2) *dogmas* not because almost all researchers actually *kept to* these principles (quite the opposite is true), but because so many authoritative voices have *professed* that these principles are the principal philosophical or methodological rationale for their theories.

II

The following simple fact that will be used to attack principle (1) draws only on the basic requirements for belief revision. Irrespective of the prior belief set K, I shall call any consistent and logically closed belief set including a sentence ϕ a *candidate revision* of K by ϕ.

Observation 1. No two distinct belief-contravening candidate revisions of a consistent and logically closed belief set by a sentence ϕ can be set-theoretically compared in terms of the sets of beliefs on which they differ with the prior belief set.[16]

This observation shows that the most straightforward idea to measure conservativity in terms of a comparison of the sets of beliefs in which original and revised theories differ fails: there are no two candidate revisions such that the symmetric difference between the one and the original belief set K is a proper subset of the symmetric difference between the other and K. If we use this criterion, then *all* (successful and closed) candidate revisions of a consistent belief set are minimally removed from it. Maxim (1) therefore cannot be a good recommendation. In one reading, it can be used to license the choice of any arbitrary (successful and closed) revision. In another reading, one which recommends the cautious strategy of taking the greatest common denominator of all minimally distant belief sets, it has as a result that all belief-contravening belief changes are amnesic.[17] Here I call a revision *amnesic* if the revised belief set consists of nothing else but the logical consequences of the sentence to be revised with; otherwise, I call it *anamnestic*.

---

[16] Proof of observation 1. Let $\neg\phi$ be in a consistent belief set K, and let K' and K'' be two distinct candidate revisions of K by ϕ. By the postulates of closure and success, we know that both K' and K'' are logically closed and contain ϕ. We make use of the following definition: for any two sets X and Y, the symmetric difference $X \Delta Y$ between X and Y is $(X \setminus Y) \cup (Y \setminus X)$. We want to show then that there is a sentence which is in $K \Delta K'$ but not in $K \Delta K''$ and there is also a sentence that is in $K \Delta K''$ but not in $K \Delta K'$. Since $K' \neq K''$ there is either a sentence in $K' - K''$ or there is a sentence in $K'' - K'$. Without loss of generality, assume the latter, and take some sentence ψ that is contained in K'' but not in K'. Then, by closure, $\neg\phi \vee \psi$ is in K and in K'' but not in K'. Hence $\neg\phi \vee \psi$ is in $K \Delta K'$ but not in $K \Delta K''$. On the other hand, again by closure, $\phi \wedge \psi$ is in K'' but neither in K nor in K' (here we also use the consistency of K). Hence $\phi \wedge \psi$ is in $K \Delta K''$ but not in $K \Delta K'$. We conclude that $K \Delta K'$ and $K \Delta K''$ are not related by subset inclusion. Quod erat demonstrandum

Observation 1 bears some similarity with David Miller's and Pavel Tichý's celebrated refutations of Karl Popper's early concept of verisimilitude; see Miller, "Popper's Qualitative Theory of Verisimilitude," *British Journal for the Philosophy of Science*, xxv (1974): 166-77; and Tichý, "On Popper's Definition of Verisimilitude," *British Journal for the Philosophy of Science*, xxv (1974): 155-60. While Popper, Miller, and Tichý are interested in the distances of (false) theories from the theory containing all and only the truths, I am interested here in the distances of (belief-contraveningly revised) candidate theories from a given prior theory.

Observation 1 seems to be a strong indication that in order to make good sense of the idea of informational economy, we need to turn to a refined description of belief states, such as the one afforded by the ordering of beliefs in terms of their epistemic entrenchment. The intuition behind the term 'entrenchment', in the sense that has been given a formal analysis in the literature on belief revision, is basically that of a relation of comparative retractability. A belief ϕ is *more entrenched* than another belief ψ if and only if the agent holds on to ϕ and gives up ψ after learning (or hypothetically assuming) that either ϕ or ψ is (or may be) false.[18]

Having obtained a rudimentary understanding of the term 'entrenchment' as it is used in the AGM tradition, let us now turn to principle (2). Like the previous observation, the following one is mathematically trivial, but conceptually striking. I call a new piece of information *moderately surprising* with respect to a belief set K if its negation is an element of K which is more than minimally entrenched in K.

Observation 2. Suppose we want to revise a belief set by a sentence ϕ and find two elements of the belief set that nonredundantly entail the negation of ϕ. Then it may well be rational, according to the standard belief-revision constructions, to restore consistency by removing the *more* entrenched and retaining the *less* entrenched belief.

---

[17] That this is so follows from Observation 2.2 in Alchourrón and Makinson, "On the Logic of Theory Change: Contraction Functions and Their Associated Revision Functions," *Theoria*, XLVIII (1982): 14-37.

[18] For official formal definitions, see Gärdenfors and Makinson, p. 89; and my "Preferential Belief Change Using Generalized Epistemic Entrenchment," *Journal of Logic, Language and Information*, I (1992): 45-78, here p. 61.

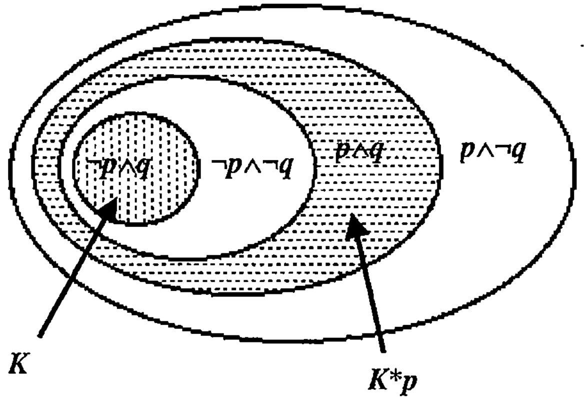


Figure 1: Example for observation 2

In fact, such a situation can *always* be identified in an anamnestic revision by a consistent and moderately surprising sentence.[19]

Principle (2) thus turns out to be plainly wrong on the "local" interpretation given to it in this observation (where reference is made to entailment sets with only two elements). Figure 1 gives a simple illustration of the situation described in observation 2, using a Grovean possible worlds representation of AGM-style belief change.[20] In the center of such a system of spheres we find the worlds satisfying all current beliefs. The entrenchment of a belief ϕ is measured by the distance of the closest $\neg\phi$-worlds from the center. The result of revising the current belief set K by a new piece of information ψ is the set $K * \psi$ of sentences satisfied by all closest ψ-worlds.

The situation is this: the incoming information is p, and two beliefs, q and $q \supset \neg p$, form a set that minimally implies the negation of p. The belief q is strictly less entrenched than the belief $q \supset \neg p$. And yet, in the revised belief set $K * p$, the agent keeps the less entrenched q and abandons the more entrenched $q \supset \neg p$—and all this according to the widely accepted AGM methodology![21]

---

[19] What *is* excluded by the definition of entrenchment advocated by Gärdenfors and Makinson, and generalized in my "Preferential Belief Change," is that the situation of observation 2 arises when the two "culprits" ψ and χ, say, *exactly* entail $\neg\phi$, that is, when $\psi \wedge \chi$ is logically equivalent with, but not logically stronger than $\neg\phi$. For then we have, by definition, $\psi < \chi$ if and only if χ is, but ψ is not, in the contraction of the belief set with respect to $\neg\phi$, and therefore χ is, but ψ is not, in the revision of the original belief set by ϕ.

[20] See Adam Grove, "Two Modellings for Theory Change," *Journal of Philosophical Logic*, XVII (1988): 157-70, and Gärdenfors, *Knowledge in Flux*, pp. 83-86, 94-97.

[21] Here is an objection to this argument from observation 2: observation 2 uses an entrenchment relation not in the GM way, but in the way in which Alchourrón and

I believe that the paradoxical appearance of observation 2 is not due to a shortcoming of the AGM recipe, but due to the fact that principle (2) expresses a wrong ideology (or, perhaps, the fact that principle (2) gives the wrong expression to an unclear idea[22]). The situation is explained in more detail in the following proof of observation 2 with which I close this section.

Let $K = Cn(\{\neg p, q\})$, where *Cn* is the operation that takes any set of sentences to the set of its logical consequences, and consider the following sequence of sentences of increasing logical strength: $\top$, $p \supset q$, $\neg p$, $\neg p \wedge q$ and $\bot$. Define the ordering $\leq$ by putting $\phi \leq \psi$ if and only if every sentence in the sequence which implies ϕ also implies ψ. This is an alternative description of the situation represented in terms of Grovean systems of spheres in figure 1.

It is easy to check that $\leq$ thus defined is an entrenchment relation in the standard sense of Gärdenfors and Makinson. Let $<$ be the

---

Makinson—"On the Logic of Theory Change: Safe Contraction," *Studia Logica*, XLIV (1985): 405-22—use hierarchies for so-called *safe contractions*. And it is no wonder that such a misapplication should lead to unexpected results. Rejoinder: theorem 4(ii) in my "On the Logic of Theory Change: More Maps between Different Kinds of Contraction Function," in Gärdenfors, ed., *Belief Revision*, pp. 122-41, shows that entrenchment relations *can* be used for constructing safe contractions and lead to exactly the same results as when applied in the "proper" GM way. But how can this be? In the example just discussed, would not a safe contraction obviously eliminate q and keep $q \supset \neg p$? No. We can read off from figure 1 that the belief $q \supset \neg p$ gets lost since it is the minimal element in another entailment set (namely, $\{p \supset q, q \supset \neg p\}$); and q—while eliminated as the minimal element of the set $\{q, q \supset \neg p\}$—is ultimately rederivable from the "safe" elements $p \supset q$ and $p \vee q$.

[22] One could try to find a more "global" formulation of the idea. An anonymous referee pointed out that the following idea of *constrained minimization* might help. Although the example illustrating observation 2 shows that entrenchment minimization per se does not serve our purpose, minimization *subject to the constraint that p be imported* makes sense, that is, minimization with respect to the relation 'q is better entrenched than $q \supset \neg p$ among revisions that import p'. Translated to the modeling at hand, this idea might mean two different things. First, reference may be made to the posterior entrenchments of the sentences in questions. This is essentially equivalent to a question about iterated belief change which is something we cannot treat here in appropriate generality (but compare section III.C below). Second, one may relativize the entrenchments by using material conditionals with antecedent p in the comparison themselves. This amounts to a very special method of iterated belief change, so-called "irrevocable belief change" (see my "Two Methods," Section 6; and Krister Segerberg, "Irrevocable Belief Revision in Dynamic Doxastic Logic," *Notre Dame Journal of Formal Logic*, XXXIX (1998): 287-306). According to this proposal, the relevant question is not whether q is more entrenched than $q \supset \neg p$, but rather whether $p \supset q$ is more entrenched than $p \supset (q \supset \neg p)$. It can be verified that this is true in our example, and that indeed the problem mentioned in observation 2 can never arise if this sort of relativization is employed. But the idea, interesting though it is, departs too much from the wording of (2) to count as a faithful formalization of this principle. It is not posterior or conditional entrenchment but plain prior entrenchment that is supposed to govern changes of belief.

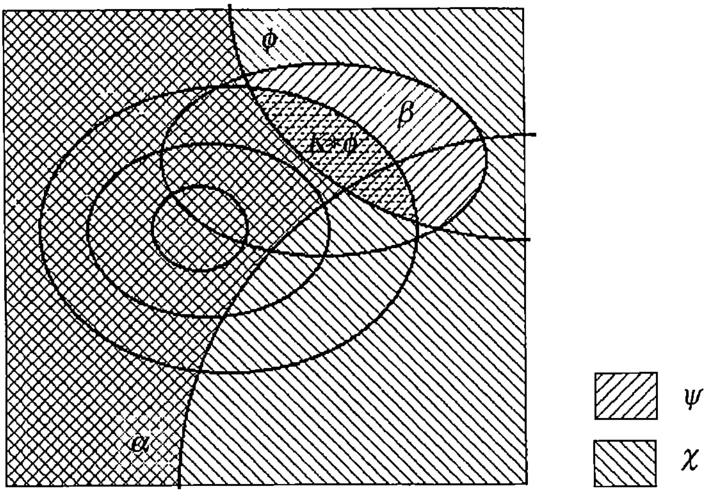


Figure 2: Illustration of the proof of the general part of observation 2

asymmetric part of $\leq$. We notice that the set of sentences that are more entrenched than $\neg p$ in terms of $<$ is $Cn(\{p \supset q\})$. Using the standard construction recipe for entrenchment-based revisions, we get that $K * p = Cn(\{\phi \in K : \neg p < \phi\} \cup \{p\})$ which equals $Cn(\{p,q\})$. Evidently, q and $q \supset \neg p$ form a nonredundant set that entails $\neg p$ ("a minimal culprit set for $\neg p$"), and q is strictly less entrenched than $q \supset \neg p$. And yet q is, but $q \supset \neg p$ is not, contained in $K * p$.

For the general part of observation 2, let ϕ be a consistent and moderately surprising sentence, and let $K * \phi$ be unequal to $Cn(\{\phi\})$. The former means that there is an α in K such that $\alpha < \neg\phi$. Take such an α, and let β be an element of $K * \phi$ which is not implied by ϕ. We need to find two beliefs ψ and χ which jointly, but not individually imply $\neg\phi$, and are such that $\psi < \chi$ but ψ is maintained while χ is abandoned in $K * \phi$. Now consider

$$\psi := (\neg\phi \wedge \alpha) \vee (\phi \wedge \beta)$$

$$\chi := \neg\phi \vee \neg\beta$$

We check that these sentences ψ and χ indeed have the desired properties (an illustration in terms of Grove spheres is given in figure 2).

First, $\psi \wedge \chi$ clearly implies $\neg\phi$. Second, ψ alone does not imply $\neg\phi$, since the sentences ϕ and β are both elements of the consistent set $K * \phi$ (here we apply the basic requirement of consistency), and their conjunction $\phi \wedge \beta$ therefore is consistent. Third, χ alone does not imply $\neg\phi$, since ϕ does not imply β and therefore $\neg\beta$ does not imply $\neg\phi$. Fourth, ψ is in $K * \phi$ since both ϕ and β are in $K * \phi$, and $K * \phi$ is closed under logical consequence, by the basic requirements of success and closure. Fifth, χ is not in $K * \phi$, since $\phi \wedge \beta$ is in $K * \phi$ and $K * \phi$ is consistent, by success and consistency. Finally, it remains to show that $\psi < \chi$.[23] From the irreflexivity of $<$, we get that not $\alpha < \alpha$. Since $\neg\phi \wedge ((\neg\phi \wedge \alpha) \vee (\phi \wedge \beta))$ implies α, we get by continuing up that not $\alpha < \neg\phi \wedge ((\neg\phi \wedge \alpha) \vee (\phi \wedge \beta))$, that is, not $\alpha < \neg\phi \wedge \psi$. From this and $\alpha < \neg\phi$, we get by conjunction up that not $\alpha < \psi$. Making use of the fact that $\alpha < \neg\phi$ once more, we deduce from this with the help of virtual connectivity (a property characteristic of standard Gärdenfors-Makinson entrenchments) that $\psi < \neg\phi$. By continuing up, we finally get $\psi < \neg\phi \vee \neg\beta$, that is, $\psi < \chi$, as desired. Quod erat demonstrandum.[24]

III

In the previous section, I attacked the almost universally held idea that informational economy alias minimum mutilation is the basis of the most promising approaches to belief change. My reason for doing so was that, on the one hand, the idea of minimal change is difficult to formulate (observation 1), and that, on the other, its application in the construction of revisions is ill-understood (observation 2).

I am now going to discuss briefly four ways of defending the minimal change idea against the above criticism. The first defense claims that I have misdirected my attack by leveling it against belief revisions rather than against the more fundamental operations of belief contraction for which even two ways of characterizing minimality seem available. The second defense counters my arguments by saying that I have been forgetful of nothing less than the core of the AGM theory, to wit, the central representation theorems linking the well-known AGM rationality postulates to the existence of orderings that are suitable for determining a reasonable standard of minimality. The third defense advances the claim that, if Alchourrón, Gärdenfors, and Makinson have not been successful in circumscribing minimality, this was so only because their notion of a doxastic state was too simplistic, and as soon as we introduce richer models that in addition to plain beliefs also encode dispositions for belief change, we shall find a natural way to make conservatism both comprehensible and defensible. The fourth defense argues that we should take up the connection of beliefs with the real world that has been lost by the usual "internalistic" approaches to belief change. If we respect truth as an aim of inquiry in our characterizations of minimal change, the disturbing results may be expected to disappear.

---

[23] In the following, I refer to the convenient axiomatization of entrenchment relations given in my "Preferential Belief Change."

[24] Notice that the proof of the general claim of observation 2 does not presume that belief revision is effected according to the traditional Gärdenfors-Makinson construction recipe for contractions. In fact, it does not depend on any particular recipe for entrenchment-based belief changes, but draws only on the logical structuring of the entrenchment relations that Gärdenfors and Makinson introduced.

I shall argue that all these attempts to save the two dogmas, though promising at first sight, are mistaken.

III.A. Contractions. The first attempt to save the idea of informational economy as a cornerstone of the traditional theories of belief change is to recall that, most of the time, Alchourrón, Gärdenfors, Makinson and their followers have written about belief *contractions*, that is, the problem how rationally to give up a belief without acquiring any new belief. Every contracted belief set is by definition a subset of the prior set, so here the problem with symmetric differences (that made it possible to prove observation 1) does not arise. Thus we have an obvious standard by which to measure differences. If a candidate contraction[25] K' is a proper subset of another candidate contraction K'', then K' is farther removed from the original set K than K'', since the set-theoretic difference $K - K'$ properly includes $K - K''$. An early suggestion of Isaac Levi's[26] is that each legitimate revision by ϕ must be decomposable into a contraction of K by $\neg\phi$, followed by an addition of ϕ and logical closing-up (this is the so-called Levi identity). In this way, distances of revisions from the prior set are indeed expressible in terms of distances of corresponding contractions from the prior set. Unfortunately, this is not intuitively plausible. If K' and K'' are two candidate contractions with respect to $\neg\phi$ and if K' is a subset of K'', then the result of adding ϕ to both of them and closing them up under logical consequences surely leads to a belief set $Cn(K' \cup \{\phi\})$ that is a subset of the belief set $Cn(K'' \cup \{\phi\})$. But is the latter closer to K than the former? There is no justification for saying so. While the latter set contains original beliefs that may have been lost in the former, it also contains novel beliefs that were not held before and can therefore be considered to be gratuitous additions. This fact is basically what observation 1 is reflecting.

Another advantage of focusing on contractions might be seen in the fact that Alchourrón, Gärdenfors, and Makinson work with the so-called postulate of recovery for contractions. This postulate says that when contracting with respect to some sentence ψ, we may withdraw beliefs only to such an extent that adding back ψ immediately after the contraction will make us recover all the original beliefs. This postulate can indeed be regarded as an explicit condition of minimum mutilation. As a general defense of minimum mutilation, however, the argument fails for three reasons. First, recovery is at best a partial encoding of informational economy, since it does not even disallow amnesic belief change.[27] Second, the information-preserving effects of recovery evaporate completely when contractions are used as intermediate steps in the construction of belief revisions with the help of the Levi identity. Third, meeting the requirement of recovery has counterintuitive consequences in many situations. This postulate has accordingly suffered from severe criticism by numerous members of the belief-revision community and cannot be considered as belonging to the core of traditional belief-revision theory any longer.

---

[25] A *candidate contraction* of a belief set K with respect to a sentence ϕ is a subset of K which is logically closed and does not contain ϕ. Here ϕ is supposed to be nontautological.

[26] "Subjunctives, Dispositions and Chances," *Synthese*, XXXIV (1977): 423-55.

I conclude that neither of the two arguments to support the idea of minimal change through a consideration of belief contractions is convincing.

III.B. Reconstructions. The success of the AGM program was not in the first place based on their putting together a list of "rationality postulates," but rather on their showing that the belief-change behavior thus axiomatized can be represented by a number of interesting and plausible explicit constructions (such as the above-mentioned revisions based on partial meets and revisions based on entrenchments). All of these constructions make use of some kind of extralogical structure of the belief state that guides the selection of most plausible or least plausible worlds or sentences featuring in the process of belief change. Typically, this selection is determined by a preference relation that is independent of the particular sentence that is to be accepted or withdrawn. The second line of defense against my attack on the notion of minimal change in belief change is based on the following argument: it is just the upshot of the many representation theorems in the literature that rational belief change can be reconstructed as belief change determined by a minimization condition with respect to some underlying doxastic preference relation.[28] Such preference relations—hidden structures of belief as it were—range alternatively over beliefs, sets of beliefs, models, or worlds.[29] And at least in the Grovean possible worlds representation, the minimization procedure has been taken to reflect a form of minimal change: from the models that satisfy the original theory, the agent passes over to the minimally remote ("closest") worlds that satisfy the input sentence.

---

[27] In terms of the Grovean possible worlds model, recovery disallows the gratuitous admission of ϕ-worlds in a contraction with respect to ϕ. It says nothing at all about the admission of $\neg\phi$-worlds—compare Maurice Pagnucco and Rott, "Severe Withdrawal (and Recovery)," *Journal of Philosophical Logic*, XXVIII (1999): 501-47.

[28] In analogy with the tradition of rational-choice theory associated with leading economists like Kenneth Arrow and Amartya Sen, one could advance the slogan "rational change is relational change." In rational-choice theory, the corresponding

A first reply to this defense is that contrary to the declarations I cited in section I, the AGM postulates for belief revision do not place *any* constraints regarding the preservation of beliefs in the case where the incoming information is inconsistent with the current theory—the interesting belief-contravening case.[30] So if there is no encoding of minimal change or informational economy in the postulates (and this is what I am claiming), it would be strange if we could conjure up such an idea through mathematical representation theorems.

Second, it needs to be supported by extra arguments, I think, that the canonical preference relations constructed in the proofs of AGM-style representation theorems are more than just technical devices and can indeed be given an interpretation that fits the desired economical meaning. Technically, it is of course possible to reverse the underlying preference orderings and select the worlds that are maximally remote ("farthest away") from the prior belief state.

Third, we have seen in observation 2 that in the case of epistemic entrenchment, the application of the stipulated orderings is not just

---

slogan is "rational choice is relational choice." Belief change can accordingly be incorporated as a subtheme into the study of the *homo oeconomicus*; see my *Making Up One's Mind: Foundations, Coherence, Nonmonotonicity* (Habilitationsschrift, Department of Philosophy, University of Konstanz, October 1996; revised version to appear with Oxford University Press under the title *Change, Choice and Inference*), and my "Logic and Choice," in Itzhak Gilboa, ed., *TARK '98: Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge* (San Francisco: Kaufmann, 1998), pp. 235-48. Gärdenfors-Makinson style entrenchment relations can be conceived as revealed preferences in precisely the sense that is common in rational choice theory.

[29] Compare Hansson, "Hidden Structures of Belief," in Fuhrmann and Rott, eds., *Logic, Action and Information* (Berlin: de Gruyter, 1995), pp. 79-100. For the now classic accounts of the 1980s, see Gärdenfors, *Knowledge in Flux*, chapters 3 and 4.

[30] The AGM postulates for belief revision include the above-mentioned "basic" requirements of success, logical closure and consistency, as well as a postulate to the effect that it is the content of a new piece of information that matters, not its syntactical formulation. Moreover, Alchourrón, Gärdenfors, and Makinson have two postulates identifying a revision by a sentence consistent with the prior belief set with the deductive closure of the union of both. These two postulates are *the only* AGM postulates for revision that lend themselves to a minimal-change interpretation. Finally, two "supplementary" AGM postulates relate revisions by varying inputs. For a more thoroughgoing interpretation of these postulates, see my "Coherence and Conservatism in the Dynamics of Belief, Part I: Finding the Right Framework," *Erkenntnis*, L (1999): 387-412.

straightforward minimization.[31] In other modelings (like partial meet contraction, safe contraction, or possible worlds models), minimization is only one step in a complex procedure of constructing revised belief sets, and its effects are at least partially neutralized in subsequent steps of this very procedure. There are tacit principles at work here, according to which a believer should respect ties in her underlying preferences and should treat equally objects that she holds in equal regard.[32] And it is again implicit in the construction of AGM-style belief changes that these principles are given priority over principles of minimal change. As a result, these constructions even license amnesic belief change.

I conclude that the argument to support the idea of minimal change through a rational reconstruction of belief revision in terms of hidden preference relations is not convincing.

III.C. Dispositions. The third line of defense of minimal change concedes that the arguments presented in section II are correct—provided that they are viewed as an attack on the particular form of early belief-change models. But the blame can be laid on an illegitimate identification of belief states with deductively closed sets of sentences, an identification that Alchourrón, Gärdenfors, and Makinson seemed to advocate themselves. On a proper understanding, so the defense continues, the formal model of a belief state should already encompass the means necessary for performing belief changes, and should therefore include something like the preference structures I discussed in the previous section. Such a move will allow us not only to devise a smooth mechanism that can easily cope with iterated changes of belief, but also to get a grip on the elusive notion of minimal change in belief-contravening revisions. If we extend AGM theory and conceive more sophisticated structures (for example, orderings or selection functions over worlds or sentences) as representations of doxastic states, we can find natural ways of defining distances between them. On this construal, the presumption underlying observation 1—namely, that in order to define minimal changes of belief states one has to compare differences between closed belief sets—is wrong.

---

[31] The official Gärdenfors-Makinson recipe for belief contractions based on entrenchments makes use of a disjunction and therefore resists an easy understanding: ψ is in the contraction of K with respect to ϕ if ψ is in K and $\phi \vee \psi$ is more entrenched than ϕ. See Gärdenfors and Makinson, *op. cit.*, p. 89.

[32] See the *principles of preference and indifference* discussed in Pagnucco and Rott. These principles tell us to go for a compromise if there are multiple optimal solutions, and are thus responsible for the fact that (2) pulls in a different direction than the uncompromising minimal change principle (1). Further relevant arguments are given by Tor Sandqvist, "On Why the Best Should Always Meet," to appear in *Economics and Philosophy*, XVI (October 2000).

Two different but ultimately equivalent ways of implementing this idea are presented in papers by Boutilier and myself,[33] where the former is based on a possible worlds modeling similar to Adam Grove's, while the latter is based on epistemic entrenchment relations. Both Boutilier and I furnish an explicit formal interpretation of minimal change. The basic idea is that in order to effect a change of the belief state represented by a certain ordering (of possible worlds or sentences), one should change in this ordering only the positions of a uniquely identifiable *minimal*[34] set of elements, just as much as is necessary in order to make the change "successful." It has turned out, however, that the possibility to come up with a truly conservative definition of (iterated) belief change has to be dearly purchased. Darwiche and Pearl were the first to make this observation by way of an intuitive counterexample the force of which is acknowledged by Boutilier.[35] A more general twist to the problem is added by the following observation.[36]

Observation 3. If doxastic states encompass revision-guiding structures (like preference orderings or selection functions), then belief-contravening revisions that obey the maxim of minimum mutilation have unacceptable consequences: they violate a requirement of temporal coherence.

I state this theorem without proof. The gist of the argument concerns iterated belief changes and the role of the recency of information. The postulate of success rules that a piece of information that is just coming in (is "most recent") is maximally appreciated and therefore invariably accepted. When the next revision takes place, however, the information just taken up (now the "second most recent" belief) turns out to be very weakly entrenched. As Darwiche and Pearl's red-bird example shows, conservative belief change in this sense can be too dismissive: he who revises a tabula rasa belief state first by 'Fred is a bird', then by 'Fred is red', and finally by 'Fred is not a bird' will end up in a belief state that does not include 'Fred is red'—and that is certainly a counterintuitive consequence.

---

[33] See Boutilier, "Revision Sequences and Nested Conditionals," in R. Bajcsy, ed., *IJCAI-93: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (San Mateo, CA: Kaufmann, 1993), pp. 519-25, and "Iterated Revision and Minimal Change of Conditional Beliefs"; and my "Coherence and Conservatism in the Dynamics of Belief, Part II: Iterated Belief Change without Dispositional Coherence," submitted to *Studia Logica*.

[34] Modulo the principle of indifference mentioned in footnote 32.

[35] See Darwiche and Pearl, "On the Logic of Iterated Belief Revision," in Ronald Fagin, ed., *TARK '94: Proceedings of the Fifth Conference on Theoretical Aspects of Reasoning About Knowledge* (Pacific Grove, CA: Kaufmann, 1994), pp. 5-23, here pp. 9 and 19; Darwiche and Pearl, "On the Logic of Iterated Belief Revision," *Artificial Intelligence*, LXXXIX (1997): 1-29, here pp. 10 and 21; and Boutilier, "Iterated Revision and Minimal Change of Conditional Beliefs," *op. cit.*, p. 296.

[36] Taken from my "Coherence and Conservatism in the Dynamics of Belief, Part II."

In another sense, conservative belief change is too tenacious. Revising the belief state that includes 'Marc is either French or Belgian' first by 'Marc is not French' and then by 'Marc is not Belgian' will preserve the initial disjunction (and even yield that Marc is French), which I think is again contrary to intuition. In conservative belief change as defined by Boutillier and myself, incoming information indeed gets always accepted but only at the lowest level possible, so in a subsequent case of conflict it has to take all the blame and is the first candidate for removal.[37] In this regard, therefore, old beliefs are treated with more respect than new beliefs. This form of belief change shows no principled stance toward the recency of information and leads to a doxastic behavior that must be called incoherent with respect to time.

In sum, it is true that more encompassing representations of belief change are desirable and even necessary, if we are to deal with the important problem of iterated revisions. The most natural suggestion how to reflect the idea of minimal change in such a framework, however, does not lead to a satisfactory solution of the belief-revision problem. The argument proposing that the idea of minimal change or informational economy can be supported through an enrichment of the notion of a belief state by structures representing doxastic dispositions is not convincing.

III.D. Truths. Except for the principle of minimum mutilation, little work has been done in the theory of belief revision to account for Quine's criteria that I mentioned in the introduction. It is particularly irritating that the "correspondence" of our beliefs with the real world—think of Quine's insistence that the empirical phenomena have to be got right—is not at all captured by the usual modelings.[38]

---

[37] The input is accepted by the agent, but, in the terminology introduced above, its negation is considered to be not even moderately surprising.

[38] Symptomatic of a certain lack of interest in *truth* among belief-change theorists is the fact that the most important book on *belief* revision has the title "*Knowledge* in Flux" (Gärdenfors, *op. cit.*) but does not attend to the truth of our alleged knowledge (nor to its justification, for that matter). The same terminological unconcern is present in the commonly used term '*epistemic* entrenchment' which should really be '*doxastic* entrenchment'—unless one does not mind disregarding more than two millennia of philosophical tradition (*vide* Plato). That belief revision today is finally getting closer to the concept of truth can be seen from the recent work of Kevin Kelly and his colleagues. See Kelly, Oliver Schulte, and Vincent Hendricks, "Reliable Belief Revision," in Maria Luisa Dalla Chiara et alia, eds., *Logic and Scientific Methods: Proceedings of the 10th International Congress of Logic, Methodology and Philos-*

More than a hundred years ago, William James formulated the main goal of belief fixation and belief change in his famous lecture on "The Will to Believe"[39]: "*We must know the truth*; and *we must avoid error*—these are our first and great commandments as would-be knowers" (*ibid.*, p. 24). As one of the founding fathers of pragmatism, James is a predecessor of Quine's as well as of Levi's, a leading philosopher of belief change, who did take over James's catchword. The charge now against formal accounts of belief change, which can also be turned against my way of capturing minimal change in principle (1) above, is that one should worry more about truth. Only true beliefs are valuable, and there is no point in preserving false beliefs. A suitably qualified version of (1) which takes into account the basic concern about truth, so the idea of the fourth defense of minimal change, would not run into the difficulties described above.

The concern with truth is to be welcomed, no doubt, but unfortunately it does not offer an escape from the predicament we have come across. First of all, observation 1 of course remains valid in a version that replaces each belief set by the set of *true* beliefs which it contains.

Corollary 4. Let K' and K'' be two belief-contravening candidate revisions of a consistent and logically closed belief set K by a sentence ϕ. If K' and K'' have different sets of true beliefs, they cannot be set-theoretically compared in terms of the true beliefs on which they differ with K.

But then, looking at symmetric differences does not seem to be what we are interested in, since in general we would not mind, but welcome *many* more true beliefs than we had prior to the revision. The dyadic notion of minimal change (referring to a relation between old and new belief set) is now dominated by the monadic notion of truth (as an aim

---

ophy of Science (Boston: Kluwer, 1997); and Kelly, "Iterated Belief Revision, Reliability, and Inductive Amnesia," *Erkenntnis*, L (1999): 11-58.

[39] In F.H. Burkhardt, F. Bowers and I.K. Skrupskelis, eds., *The Will to Believe and Other Essays in Popular Philosophy: The Works of William James, Volume 6* (Cambridge: Harvard, 1979), pp. 13-33. Incidentally, one decade later, in 1907, James takes the "observable process...by which any individual settles into *new opinions*" to show that "in this matter of belief we are all extreme conservatives" and that we invariably aim at a "minimum of disturbance" (or "minimum of modification", or "minimum of jolt") of the older stock of beliefs—*Pragmatism: The Works of William James, Volume 1*, F. Bowers and I.K. Skrupskelis, eds. (Cambridge: Harvard, 1975), pp. 34-35. James says that "loyalty" to older truths is not only the first, but often the only principle for belief change. This is formulated as a descriptive statement about empirical believers, and as such it is part of the pragmatist folklore; see Peirce, p. 66: "belief...is a calm and satisfactory state which we do not wish to avoid, or to change to a belief in anything else. On the contrary, we cling tenaciously, not merely to believing, but to believing just what we do believe." In contrast to Peirce, James in the passages just cited quite evidently approves of "extreme" conservatism as a normative doctrine.

of the posterior belief set). What we want to make sure is really that we minimize the loss of true beliefs. This does not lead to a satisfactory result either, however, as is borne out by the following observation.

Observation 5. Let K' be a belief-contravening candidate revision of a logically closed belief set K by a sentence ϕ. Then K' minimizes the set of true beliefs lost from K (among all other candidate revisions of K by ϕ) only if K' is *opinionated* in the sense that it contains either ψ or $\neg\psi$, for every sentence ψ.[40]

It is clear that the commitment to opinionated theories is undesirable, especially since the prior theory may well be undecided about countless contingent matters. So minimizing the loss of true beliefs cannot be the only criterion relevant for our purposes. It is worth pointing out explicitly that Alchourrón, Gärdenfors, and Makinson themselves have always been decidedly against the unbridled use of informational economy as it manifests itself in the inflationary behavior of so-called (maxi-)choice contractions and revisions of logically closed belief sets.[41]

I conclude that the argument to support the idea of minimal change through a relativization to true beliefs does not succeed.

As a little digression, I take down the following, perhaps even more surprising observation.

Observation 6. No belief-contravening candidate revision that does not contain every truth strictly enlarges the set of true beliefs. In particular, even if the prior theory has been false while the new piece of information as well as everything else in the posterior theory is true, the set of truths is not strictly increased.[42]

---

[40] Observation 5 stands to observation 3.2 of Alchourrón and Makinson, "On the Logic of Theory Change: Contraction Functions and Their Associated Revision Functions," essentially as corollary 4 stands to observation 1. For the sake of transparency, I give an independent proof of observation 5. Let the belief set K be consistent and logically closed, let $\neg\phi$ in K, and let K' be a candidate revision by ϕ which is not opinionated. We show that K' does not minimize the set of lost true beliefs from K. Assume first that ϕ is true. Then the set T of all true sentences is a candidate revision containing ϕ. Take a true sentence ψ that is not contained in K'; there is such a ψ since K' is not opinionated. Then, by logical closure, $\neg\phi \vee \psi$ is a true sentence that was contained in the prior belief set K and is lost in the candidate belief set K'. But $\neg\phi \vee \psi$ is in T, and T loses no true belief from K at all, so T actually loses less true beliefs from K than K'. Thus K' does not minimize the set of lost true beliefs from K. Assume secondly that ϕ is false. Then $\neg\phi$ is true. Extend K' to an opinionated set K'', and take some ψ from $K'' - K'$. Then, by logical closure, $\neg\phi \vee \psi$ is a true sentence from K which is lost in K' but not in K''. Since K' is a subset of K'', K'' loses no true sentence from K that K' does not lose. So K'' actually loses less true beliefs from K than K'. Thus, K' does not minimize the set of lost true beliefs from K. Quod erat demonstrandum

[41] Compare the discussion in Alchourrón and Makinson, pp. 20-21, and in Makinson, pp. 356-59.

It turns out that except for God (and perhaps various demons who can jump to the whole truth in a single leap) there is no one-way road to the truth. For every newly acquired truth, we have to pay a price by trading in other truths. We may hope that these truths are less important or fundamental than the ones we acquire, but this is an issue that cannot be settled in an a priori manner.

IV

Getting clear about foundations is not just of theoretical interest, but has a practical effect as well: we expect to learn in which direction we are to head if we wish to improve our current theories of belief change, as well as their application in knowledge-based systems. The main point here is that the theories developed by Alchourrón, Gärdenfors, and Makinson and their followers are *not* oriented toward the principle of informational economy, and I have found no reason why they should. I have called the two versions of this principle dogmas because many researchers (including the present author) have believed in it and recited it time and again, without actually keeping to it when building their theories.

We have seen that contrary to many pronouncements of belief-change theorists, it is even difficult to spell out what is meant exactly by the idea of informational economy or minimum mutilation, as long as belief change is to satisfy the basic requirements. Theories of belief change in the tradition of Alchourrón, Gärdenfors, and Makinson do not align themselves easily with the idea of minimal change, and they are by no means centered on this idea. I have tried to show that two of the most natural ways of fleshing out the idea do not achieve what they are widely supposed to achieve, and that four attempts to defend minimal change against my attack fail.

I should like to stress, however, that I have only been concerned with principles that are supposed to describe what motivates an important class of existing normative theories about belief change. I have refrained from putting forward any normative thesis about belief change myself, neither against nor in favor of conservatism. This is an entirely different undertaking.[43]

---

[42] Proof of observation 6. Let $\neg\phi$ be in the belief set K, and let K' be a candidate revision of K by ϕ. By the postulates of closure and success, we know that K' is logically closed and contains ϕ. Now we let K_t and K'_t be the set of true beliefs in K and K', respectively. What we want to show that K'_t is not a strict superset of K_t, that is, that there is a true sentence which is in K but is lost in K'. Take some true sentence ψ which is not in K'. Such a sentence exists since K' is assumed not to be omniscient. Now consider $\neg\phi \vee \psi$. This sentence is true, since ψ is true, and it is in K, since $\neg\phi$ is in K and K is logically closed. $\neg\phi \vee \psi$ is not in K', however, since ϕ is in K' and ψ is not in K'. Thus, we have found a truth that has been lost. The second part of observation 6 follows from the first. Quod erat demonstrandum.

Yet another point deserves attention. Little work—perhaps no work at all—has been done that reflects Quine's criteria (a) and (b) in the theory of belief revision. In his joint book with Joseph S. Ullian, *The Web of Belief*,[44] Quine added more virtues that good theories should have: modesty, generality, refutability, and precision. Again, belief revision as studied so far has little to offer that would reflect these intuitive desiderata. Except for the issue of conservatism, Quine's list is a list for theory *choice* rather than for theory *change*, because it lists properties that a good posterior theory should have, independently of any properties of the prior theory. It is a strange coincidence that the philosophy of science has focused on monadic (nonrelational) features of theory choice, while philosophical logic has emphasized dyadic (relational) features of theory change. I believe that it is time for researchers in both fields to overcome this separation and work together on a more comprehensive picture.

Here, however, I first of all hope to draw attention to the fact that it is not appropriate to focus exclusively on the idea of informational economy even when talking about nothing but the restricted, AGM-style modelings of belief change. There are various coherence criteria that find expression in formal "rationality postulates": inferential coherence (consistency and closure), dispositional coherence (a kind of semantic representability that can be encoded syntactically in so-called "supplementary" rationality postulates—which are what I would call the hallmark of the AGM theory), as well as temporal coherence (a principled appreciation of the recency of information in iterated belief change). These criteria substantially restrict the dominion of informational economy in formal belief-change theories. Having understood this, we may expect that the picture will change again enormously when we set out to live up to the Quinean epistemological virtues. It is time to liberate our minds from the constrictions of informational economy and give conscious admittance to other norms for our ethics of belief.

HANS ROTT

University of Regensburg

---

[43] Compare Gilbert Harman, *Change in View* (Cambridge: MIT, 1986) and "Rationality," in E.E. Smith and D.N. Osherson, eds., *Thinking: Invitation to Cognitive Science*, Volume III (Cambridge: MIT, 1995), pp. 175-211; reprinted in Harman, *Reasoning, Meaning and Mind* (New York: Oxford, 1999), pp. 9-45. A good critical discussion of conservatism in contemporary normative epistemology is given by David Christensen, "Conservatism in Epistemology," *Noûs*, XXVIII (1994): 69-89.

[44] New York: Random House, 1978, second edition.

BOOK REVIEWS

Causality and Explanation. WESLEY C. SALMON. New York: Oxford University Press, 1999. 434 p. Cloth $65.00, paper $29.95.

It is true of very few philosophers that their work has changed a subject's direction. Wesley Salmon's did in the early 1970s, when he convincingly argued that Carl Hempel's deductive-nomological and inductive-statistical models of explanation were not just incorrect but irreparably defective. In the course of developing his statistical-relevance model, Salmon realized that an account of causality acceptable to empiricists was needed to underpin explanations and that the account must encompass both deterministic and probabilistic situations. His research on causation led to the theory of markable causal processes which served as the basis for the ontic account of explanation developed in his 1984 book, *Scientific Explanation and the Causal Structure of the World.*[1] Always open to reasoned change, Salmon re-oriented his approach once again in the 1990s, developing a conserved-quantity approach to causal transmission and interaction along lines initially suggested by David Fair[2] and more adequately developed by Phil Dowe.[3] The intellectual passage from breakthrough to almost current condition is chronicled in this excellent collection, with much else besides.[4] Although it includes seven previously unpublished papers, the core of Salmon's work is contained in the nineteen reprinted articles across which his philosophical development in these areas can be traced. Taken both individually and collectively, this is a richly rewarding selection of articles around which an excellent seminar could be constructed.

The statistical-relevance model undercut the nomic-expectability basis of Hempel's model by showing that events with low probabilities could be explained. This was done by citing factors that made the events either more or less probable. (Whether it should have allowed explanations in terms of factors that lowered the probability is a famously divisive issue, but we can here pass over that point.) To

---

[1] Princeton: University Press, 1984.

[2] "Causation and the Flow of Energy," *Erkenntnis*, XIV (1979): 219-50.

[3] "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science*, LIX (1992): 195-216.

[4] The collection does not include some of Salmon's most recent papers. One important omitted essay is his "Causality and Explanation: A Reply to Two Critiques," *Philosophy of Science*, LXIV (1997): 461-77.

0022-362X/00/9709/523–27

explain an event, said Salmon, was to assign it to the broadest homogenous reference class, that class being arrived at by successive specifications of statistically relevant factors. This was appealing, and the idea generated a flood of literature. Yet the task of satisfactorily completing Hans Reichenbach's project of developing a purely probabilistic theory of causation proved elusive for Salmon. Faced with the problem of distinguishing between mere correlations and genuine causal relations—*the* problem for all accounts of causation—Salmon decided to view statistical-relevance relations as simply the evidential starting point of explanations, a basis that itself had to be explained. This was done in terms of an ontology of spatially continuous processes and of interactions between those processes. But the criterion of a genuine causal process, as opposed to a pseudo-process, was its ability to have its structure altered after an interaction. This criterion seemed to have a counterfactual claim embedded in it—what would happen if a mark were to be introduced—and conscientious empiricists such as Salmon want no part of irreducible counterfactual claims. So causation was instead taken to consist in the transfer of conserved quantities: "causal processes transmit conserved quantities; and by virtue of this fact, they are causal..." (253).[5]

How satisfactory is this? For one thing, it implicitly commits us to physicalism in a way that might give some pause. There are, as far as I know, no nontrivial conservation laws peculiar to the social sciences and few, if any, in biology. This means that all anthropological and sociological causation, for example, must be accounted for by the transfer of mass-energy, linear and angular momentum, and so on. And so all explanations in those sciences must ultimately be given in terms of physical causation. Perhaps this is, in the end, what physicalists must believe, and it presents in a stark form the kind of reductionist ontology that produces the various problems of mental causation. Yet here the problem reaches far down into the chemical and biological realms. In contrast, analyses in terms of sufficient or necessary conditions, or those in terms of statistical relations, have no difficulty in dealing with social or economic causation—they are subject-independent accounts. Even markable processes can, in principle, occur within any material subject matter. Marked bills, distorted states of consciousness, and changes in social structure are all plausible examples within, respectively, economics, psychology, and sociology. The conserved quantity approach thus eliminates many of the examples that made the statistical-relevance model initially plausible. We explain delinquency in terms of broken homes, unemployed fathers, depressed economies, and low levels of education, not in terms of conserved quantities. Of course, mass-energy conservation holds in the social realm as well, but it does not seem to be the relevant variable.

---

[5] Salmon's position on whether to require conserved or relativistically invariant quantities has shifted back and forth over time. His current view is to demand that they be conserved.

The emphasis on the conservation of mass-energy, momentum, and charge might lead one to misconstrue Salmon's approach to causation because he sometimes puts his position in terms of the transmission of causal influence. (See, for example, the title of essay 12; also 224, 253, 289.) Yet not all conservation laws obviously involve the kind of "active powers" that go with this way of talking. Some conservation laws seem to involve only constraint satisfaction. Take baryon number, for example. This is a conserved quantity whenever the closed system considered contains baryons (examples of which are protons and neutrons).[6] Although there is an obvious sense in which something is exchanged during baryon interactions, no causal influence is transmitted. Rather, this sort of situation seems to exemplify a structural law of the kind that played a central role in Salmon's earlier development of interactive forks. (See essay 18.)

The important distinction for Salmon is, instead, the one between causal and pseudo-processes. The former transmit conserved quantities, the latter do not. This clarification is important because more than one quantity can be conserved in a given interaction—baryon number and mass-energy, say. If it were the transmission of causal influence that was central, we would have the analog of an overdetermination problem for the conserved quantities view. As it is, either quantity will do to make the processes and interaction causal.

There is a more general way in which Salmon's work has changed the direction of the field or, more precisely, has reopened an option that had long been neglected. The change in direction was in moving from what he has dubbed the *epistemic* approach to explanation to the *ontic* approach. We can look at this switch in various ways, but its most important consequence is lessening the emphasis on the role of language and logic in understanding explanation and causation. One adopts instead a realist ontology which uses processes, interactions, and conserved quantities. It is no longer the representation that is of primary importance but the world itself. This move is to be applauded: together with Salmon's at-at account of processes, it shifts the focus from reconstructions within logical space, which are too unconstrained, to the space of physical entities and their objective relations. Here the world itself imposes constraints upon what can and cannot occur. Moves that can be made in logical space, such as forming arbitrary disjunctive predicates, are harder to justify when physical counterparts are insisted upon.

---

[6] There is some reason to think that baryon conservation might be violated in physically extreme cases such as proton decay, but the evidence is at present inconclusive.

Traditional analyses of causation in terms of sufficient conditions, necessary conditions, or probabilistic-relevance relations are intended to be necessarily true—either it is impossible for there to be a case of causation that does not satisfy a correct analysis or, given the concepts that we have, no counterexample to a proper analysis could be intelligibly formulated by us. Salmon's approach seems to be interestingly different because, although he couches his account in terms of definitions (essay 16), it seems intended to apply only to our world. Rather than provide us with a conceptual analysis of what we mean by "causation," the intent is to illuminate how the physical world works at the causal level. The core contrast between pseudo-processes and causal processes is drawn from a scientific theory (special relativity). This raises controversial questions about the degrees of empirical and conceptual content in such theories, but Salmon's account is clearly not the result solely of an a priori analysis. Furthermore, he explicitly rejects the need for conservation *laws* ("a conserved quantity is a quantity that does not change...*it makes no difference whether or not that true generalization is lawful; only its truth is at stake*" (259)).

Most of us will agree that conservation generalizations are only contingently true. So we might ask: If there were exceptions to our conservation generalizations, would there be no causal interactions and no causal processes? To answer "Yes" to this question seems unsatisfactory. It is a difficult answer to accept because of arguments that also count against regularity theories of causation. For example, if the generalization failed only occasionally in remote regions of our universe, why should that affect the causal status of local interactions that do conserve the quantity? A more direct reason for uneasiness is this: If the world worked differently than it apparently does, perhaps because there were widespread violations of conservation principles, could we still not use the term 'cause' in exactly the way that we now do and still understand the causal workings of the world? Perhaps there is a sense in which the question we asked about what would happen if the physical world were different requires no answer in the ontic approach. All that matters is what makes for causation in the physical world, for that is all the world there is. But if so, this places Salmon's account of causation into the category of scientifically based, contingent a posteriori hypotheses and another part of metaphysics has been methodologically transformed.

These are difficult issues to evaluate conclusively, and the fact that they arise within Salmon's position is an indication of its deep philosophical importance. This collection is general philosophy of science at its best—original, honest, thought-provoking, and exceptionally clear. It is in every way a fitting successor to Hempel's *Aspects of Scientific Explanation.*

PAUL HUMPHREYS

University of Virginia

NEW BOOKS: ANTHOLOGIES

AXTELL, GUY, ed.: *Knowledge, Belief, and Character: Readings in Virtue Epistemology.* Lanham, MD: Rowman and Littlefield, 2000. xxix, 224 p. Cloth $64.00, paper $21.95.

BEISTEGUI, MIGUEL DE, and SIMON SPARKS, eds.: *Philosophy and Tragedy.* New York: Routledge, 2000. ix, 246 p. Cloth $75.00, paper $22.99.

DASTON, LORRAINE, ed.: *Biographies of Scientific Objects.* Chicago: University Press, 2000. ix, 307 p. Cloth $55.00, paper $19.00.

GOWANS, CHRISTOPHER, ed.: *Moral Disagreements: Classics and Contemporary Readings.* New York: Routledge, 2000. xii, 265 p. Paper $25.99.

HIGGINBOTHAM, JAMES, FABIO PIANESI, and ACHILLE C. VARZI, eds.: *Speaking of Events.* New York: Oxford, 2000. vii, 295 p. Cloth $55.00, paper $24.95.

HORDER, JEREMY, ed.: *Oxford Essays in Jurisprudence: Fourth Series.* New York: Oxford, 2000. vii, 270 p. Cloth $74.00.

HUMPHREY, NICHOLAS, et alia: *How to Solve the Mind-Body Problem.* Bowling Green, OH: Imprint Academic, 2000. 112 p. Paper $12.95.

KATZ, LEONARD D., ed.: *Evolutionary Origins of Morality: Cross-Disciplinary Perspectives.* Bowling Green, OH: Imprint Academic, 2000. xvi, 352 p. Paper $24.95.

LEVINE, MICHAEL P., ed.: *The Analytic Freud: Philosophy and Psychoanalysis.* New York: Routledge, 2000. xii, 320 p. Cloth $85.00, paper $27.99.

MACHAMER, PETER, MARCELLO PERA, and ARSTIDES BALTAS, eds.: *Scientific Controversies: Philosophical and Historical Perspectives.* New York: Oxford, 2000. x, 278 p.

MCCRACKEN, C.J., and I.C. TIPTON, eds.: *Berkeley's* Principles and Dialogues: *Background Source Materials.* New York: Cambridge, 2000. x, 300 p. Cloth $59.95, paper $22.95.

MISCEVIC, NENAD, ed.: *Nationalism and Ethnic Conflict: Philosophical Perspectives.* Chicago: Open Court, 2000. vi, 331 p. Cloth $39.95.

MULLER, JOHN, and JOSEPH BRENT, eds.: *Peirce, Semiotics, and Psychoanalysis.* Baltimore: Johns Hopkins, 2000. xi, 184 p. Cloth $38.00.

NÚÑEZ, RAFAEL, and WALTER J. FREEMAN, eds.: *Reclaiming Cognition: The Primacy of Action, Intention, and Emotion.* Bowling Green, OH: Imprint Academic, 2000. xix, 284 p. Paper $24.95.

SEDGWICK, SALLY, ed.: *The Reception of Kant's Critical Philosophy: Fichte, Schelling, and Hegel.* New York: Cambridge, 2000. xi, 338 p. Cloth $59.95.

STERN, ROBERT, ed.: *Transcendental Arguments: Problems and Prospects.* New York: Oxford, 1999. x, 327 p. Cloth $65.00.

TURNER, STEPHEN, ed.: *The Cambridge Companion to Weber.* New York: Cambridge, 2000. xx, 288 p. Cloth $57.95, paper $19.95.

VAN DEN BELD, TON, ed.: *Moral Responsibility and Ontology.* Boston: Kluwer, 2000. vi, 276 p. Cloth $125.00.

Philosophy from Routledge

The Turn to Ethics

Edited by Marjorie Garber, Beatrice Hanssen, and Rebecca L. Walkowitz

The contributors to this volume bring the particularities of their disciplines and training to a vital complex of questions involving ethics.

Critique of Violence

Between Poststructuralism and Critical Theory

Beatrice Hanssen

"...one of the most sophisticated theorists of her generation...[her work] casts much light and enables us to force new forms of solidarity."

—Cornel West

Desiring Whiteness

A Lacanian Analysis of Race

Kalpana Seshadri-Crooks

Through close readings of literary and film texts, Seshadri-Crooks demonstrates that race is a system of differences organized around a privileged term: Whiteness.

Hispanics/Latinos in the United States

Ethnicity, Race, and Rights

Edited by Jorge J.E. Gracia and Pablo De Grieff

"...what unifies these essays is not their gregariousness but the rigor, passion and independence of thought they emanate, all of which are pushing the young discipline to unexpected heights."

— Ilan Stavans, author of *The Hispanic Condition*

Moral Disagreements

Classic and Contemporary Readings

Edited by Christopher W. Gowans

"...rich in examples and philosophical strategies."

—Michael Krausz, Bryn Mawr College

Philosophy and Tragedy

Edited by Miguel de Beistegui and Simon Sparks

Eleven new essays by internationally renowned philosophers clearly show how time and again, major thinkers have returned to tragedy in many of their key works.

Philosophy of Law

An Introduction

Mark Tebbit

Tebbit eases readers into the most fundamental theories of law and links them to the most prevalent philosophical questions encountered in legal studies.

Relating Narratives

Storytelling and Selfhood

Adriana Cavarero

"...one of the most brilliant, nuanced, inspiring, and erudite studies in philosophy and literature to be produced by a feminist scholar in recent years."

— Judith Butler

Self-Knowledge and the Self

David A. Jopling

"...the most sustained and serious contemporary philosophical reflection on the Delphic injunction "Know thyself" of which I am aware..."

—William L. McBride, author of *Social and Political Philosophy*

29 West 35th Street, New York, NY 10001-2299

www.routledge-ny.com

At Bookstores or Call 1-800-634-7064

Susan HAACK

Now in Paper!

Manifesto of a Passionate Moderate

Unfashionable Essays

"One guidebook I would recommend for a tour through the labyrinths of modern skepticism." —Mary Lefkowitz, *The New York Times Book Review*

"[T]his new book is a gem. . . . *Manifesto of a Passionate Moderate* is an important book. It is indeed refreshing to find a philosopher in the pragmatist tradition addressing the tendencies in society that are most in need of philosophical clarification in a way that is accessible to a general audience."—Robert Talisse, *Society for the Advancement of American Philosophy Newsletter*

Paper $13.00

The University of Chicago Press 5801 South Ellis Avenue, Chicago, IL 60637
www.press.uchicago.edu

The Journal of Philosophy

subscriptions - 12 issues 2000

Students	$20.00
Individuals	$35.00
Libraries and Institutions	$75.00
Postage outside the U. S.	$15.00

Payments only in U. S. currency on a U. S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first-come, first-served basis.

Published monthly as of January 1977; printed and typeset by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. Telephone: (212) 866-1742. Fax: (212) 932-3721. All manuscript inquiries should be sent to Michael Kelly, Managing Editor, at the above address.

NEW FROM INDIANA!

THE GOOD, THE BAD, AND THE BEAUTIFUL

Discourse about Values in Yoruba Culture

Barry Hallen

"Hallen renders a great service to African philosophy." —Kwasi Wiredu

Reveals everyday language as the key to understanding morals and ethics in Yoruba culture.

January 2001 256 pp
cloth $39.95 paper $17.95

THE OTHER HUSSERL

The Horizons of Transcendental Phenomenology

Donn Welton

"Welton has an incredible grasp of the whole of Husserl's work . . . insightful and detailed, yet at the same time readable and accessible." —David Carr

An original and comprehensive reconstruction of Husserl's phenomenological method.

Studies in Continental Thought
December 2000 512 pp
cloth $49.95

AFRICAN PHILOSOPHY AS CULTURAL INQUIRY

Edited by Ivan Karp and D. A. Masolo

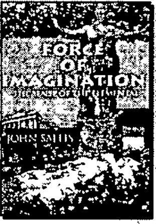
Assesses the direction and impact of African philosophy as well as its future role.

African Systems of Thought
Published in association with the International African Institute, London
December 2000 272 pp, 10 b&w photos
cloth $39.95 paper $18 95

FORCE OF IMAGINATION

The Sense of the Elemental

John Sallis



"... a powerful, original approach to what others call 'ecology' but what Sallis shows to be a question of the status of the earth in philosophical thinking at this historical moment." —Edward S. Casey

Studies in Continental Thought
October 2000 272 pp
cloth $39.95 paper $24.95

Now in a new edition!

REMEMBERING

A Phenomenological Study, Second Edition

Edward S. Casey

A *Choice* Outstanding Academic Book

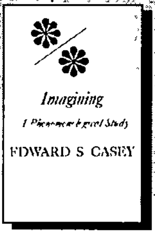
"An excellent book that provides an in-depth phenomenological and philosophical study of memory." —Choice

"... a stunning revelation of the pervasiveness of memory in our lives." —Contemporary Psychology

Studies in Continental Thought
384 pp
cloth $49.95 paper $22.95

Now in a new edition!

IMAGINING



A Phenomenological Study, Second Edition

Edward S. Casey

"This scrupulous, lucid study is destined to become a touchstone for all future writings on imagination." —Library Journal

Studies in Continental Thought
256 pp
cloth $39 95 paper $18.95

CHARLES S. PEIRCE'S PHILOSOPHY OF SIGNS

Essays in Comparative Semiotics

Gérard Deledalle

"This is essential reading for anyone who wants to find common ground between the best of American semiotics and better-known European theories. Deledalle has done more than anyone else to introduce Peirce to European audiences, and now he sends Peirce home with some new flare." —Nathan Houser, Director, Peirce Edition Project

December 2000 304 pp
cloth $39.95

at bookstores

INDIANA UNIVERSITY PRESS

Orders: 1-800-842-6796
iupress.indiana.edu

Best in Philosophy

Kant and the Fate of Autonomy
Problems in the Appropriation of the Critical Philosophy
Karl Ameriks
Modern European Philosophy

0-521-78101-9	Hardback	$54.95
0-521-78614-2	Paperback	$19.95

Phenomenal Consciousness
A Naturalistic Theory
Peter Carruthers

0-521-78173-6	Hardback	$59.95

Physical Causation
Phil Dowe
Cambridge Studies in Probability, Induction and Decision Theory

0-521-78049-7	Hardback	$54.95

The Jurisprudential Foundations of Corporate and Commercial Law
Jody S. Kraus *and* ***Steven D. Walt,*** *Editors*
Cambridge Studies in Philosophy and Law

0-521-59157-0	Hardback	$54.95

Complicity
Ethics and Law for a Collective Age
Christopher Kutz
Cambridge Studies in Philosophy and Law

0-521-59452-9	Hardback	$59.95

Law and Market Economy
Reinterpreting the Values of Law and Economics
Robin Paul Malloy

0-521-78214-7	Hardback	$64.95
0-521-78731-9	Paperback	$22.95

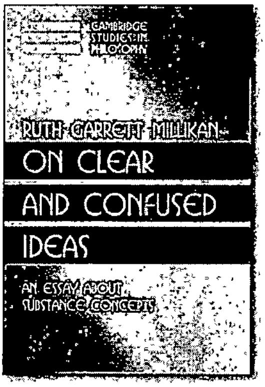
Community, Solidarity and Belonging
Levels of Community and their Normative Significance
Andrew Mason

0-521-63129-7	Hardback	$59.95
0-521-63728-7	Paperback	$21.95

from Cambridge

Plato and his Predecessors
The Dramatisation of Reason
Mary Margaret McCabe
0-521-65306-1 Hardback $59.95

On Clear and Confused Ideas
An Essay about Substance Concepts
Ruth Garrett Millikan
Cambridge Studies in Philosophy
0-521-62386-3 Hardback $59.95
0-521-62553-X Paperback $21.95



CAMBRIDGE STUDIES IN PHILOSOPHY
RUTH GARRETT MILLIKAN
ON CLEAR AND CONFUSED IDEAS
AN ESSAY ABOUT SUBSTANCE CONCEPTS

Myth and Philosophy from the Presocratics to Plato
Kathryn Morgan
0-521-62180-1 Hardback $64.95

Economics and Language
Ariel Rubinstein
Churchill Lectures in Economics
0-521-59306-9 Hardback $42.95
0-521-78990-7 Paperback $15.95

Can a Darwinian be a Christian?
The Relationship Between Science and Religion
Michael Ruse
0-521-63144-0 Hardback $24.95

Kant's Early Critics
The Empiricist Critique of the Theoretical Philosophy
Brigitte Sassen, *Editor*
Cambridge Edition of the Works of Immanuel Kant
0-521-78167-1 Hardback $54.95

Nietzsche's Postmoralism
Essays on Nietzsche's Prelude to Philosophy's Future
Richard Schacht, *Editor*
0-521-64085-7 Hardback $54.95

Available in bookstores or from
CAMBRIDGE
UNIVERSITY PRESS

40 West 20th St., N.Y., NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cup.org
AmEx/MasterCard/VISA accepted.
Prices subject to change.

Forthcoming!

HEGEL'S DIALECTICAL LOGIC

ERMANNO BENCIVENGA

This clear, accessible account of Hegelian logic offers comprehensive coverage of Hegel's important works. Bencivenga makes a case for Hegelian logic's enormous seductiveness, its surprising presence in the collective consciousness, and the dangers associated therewith.
October 2000 176 pp. $35.00

OF MINDS AND MOLECULES

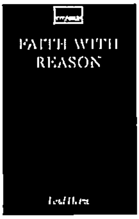
New Philosophical Perspectives on Chemistry

Edited by NALINI BHUSHAN and STUART ROSENFELD

This is the first anthology devoted exclusively to work in the philosophy of chemistry. The essays adopt distinctive philosophical perspectives on chemistry and collectively offer both a conceptualization of and a justification for this emerging field.
2000 320 pp. $55.00

FAITH WITH REASON

PAUL HELM



Helm investigates what religious faith is and what makes it reasonable. Helm argues that the reasonableness of faith depends not only on beliefs about the world but also on beliefs about oneself and on what one is willing to trust.
2000 196 pp. $39.95

HUME'S ABJECT FAILURE

The Argument Against Miracles

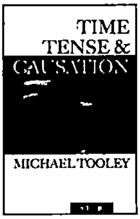
JOHN EARMAN

This vital study offers a new interpretation of Hume's famous "Of Miracles," which notoriously argues against the possibility of miracles. By situating Hume's argument in the context of the eighteenth-century debate on miracles, Earman shows Hume's argument to be largely unoriginal and chiefly without merit where it is original.
2000 256 pp. paper $24.95 cloth $49.95

New in paperback!

TIME, TENSE, AND CAUSATION

MICHAEL TOOLEY



"Tooley's provocative book makes an extremely valuable contribution to the literature on the philosophy of time and is mandatory reading for anyone working in the area."—*Mind.*
1997 (paper 2000) 416 pp.
paper $24.95 cloth $45.00

THE ANCIENT QUARREL BETWEEN PHILOSOPHY AND POETRY REVISITED

Plato and the Greek Literary Tradition

SUSAN B. LEVIN

In this study, Levin argues that Plato takes the Greek literary tradition as a key locus of ideas with respect to linguistic issues in the *Cratylus*, *Phaedo*, and the *Republic*.
2000 208 pp. $39.95

New in paperback!

NATURALISM IN MATHEMATICS

PENELOPE MADDY

"[Maddy] writes in a clear, forthright and challenging style. Her book is eminently readable, instructive, and thought-provoking."—*The Journal of Symbolic Logic.*
1998 (paper 2000) 264 pp.
paper $19.95 cloth $45.00

THE CHARACTER OF LOGIC IN INDIA

BIMAL KRISHNA MATILAL

Edited by JONARDON GANERI and HEERAMAN TIWARI

In this book, Matilal brings together two decades of thought about the origins, development, and nature of Indian logical theory. His distinctive treatment of the philosophical issues allows our understanding of classical Indian writers to be enriched by concepts drawn from contemporary analytical theory.
2000 192 pp. $12.95

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our website at **www.oup.com**

New in paperback!

POINTS OF VIEW

A.W. MOORE

Moore argues that it is possible to think about the world from no point of view. He explores issues such as the nature of persons, the subject-matter of mathematics, realism and anti-realism, value, and God. The result is a powerful critique of our own finitude.

1997 (paper 2000) 328 pp.
paper $19.95 cloth $35.00

New in paperback!

SEX AND SOCIAL JUSTICE

MARTHA C. NUSSBAUM

"*Sex and Social Justice* is highly readable, and very engaging. It is elegantly written and carefully argued." —Alan Ryan, *The New York Times Book Review.*

1999 (paper 2000)
448 pp. paper $15.95
cloth $35.00

THEMES IN HUME

The Self, the Will, Religion

TERENCE PENELHUM

Since the 1950s, Penelhum has been a leading contributor to studies on the thought of David Hume. In this collection, he presents a selection of the best of his essays on Hume.

2000 296 pp. $55.00

GOD AND GOODNESS

HUGH RICE

Rice explains why belief in God need not be seen as a strange or irrational belief, but can be a natural extension of our ordinary ways of thinking. He suggests that we should think of God in an abstract way, and he offers a satisfying account of the relationship between God and goodness.

2000 156 pp. $29.95

OXFORD STUDIES IN ANCIENT PHILOSOPHY

Volume XVIII

Edited by DAVID SEDLEY

This is an annual volume of original articles on all aspects of ancient philosophy. This year's contributions range over a thousand years of philosophy, from the Presocratics to Boethius. Particularly prominent in the volume are Aristotle and the Stoics.

2000 256 pp. paper $35.00 cloth $60.00

New in paperback!

SEXTUS EMPIRICUS: AGAINST THE ETHICISTS

(Adversus Mathematicos XI)

Translated with an introduction and commentary by RICHARD BETT

"Bett's careful translation...ought to be the main English translation consulted....Judicious and learned... Bett's work is a distinguished addition to the literature on skepticism."—*Bryn Mawr Classical Review.*

(Clarendon Later Ancient Philosophers)

1997 (paper 2000) 336 pp.
paper $24.95 cloth $65.00

EMOTION AND PEACE OF MIND

From Stoic Agitation to Christian Temptation

RICHARD SORABJI

Sorabji presents a ground-breaking study of ancient Greek views of the emotions and their influence on subsequent theories and attitudes. Focussing on the Stoics, Sorabji also draws on a vast range of texts to give a rich historical survey of how Western thinking about the emotions developed.

2000 440 pp. $45.00

TRUTH, RATIONALITY, AND PRAGMATISM

Themes from Peirce

CHRISTOPHER HOOKWAY

Hookway presents a series of studies of themes from the work of Charles S. Peirce (1839-1913). These themes center on the question of how we are to investigate the world rationally. Hookway shows how Peirce's ideas about this continue to play an important role in contemporary philosophy.

2000 280 pp. $45.00

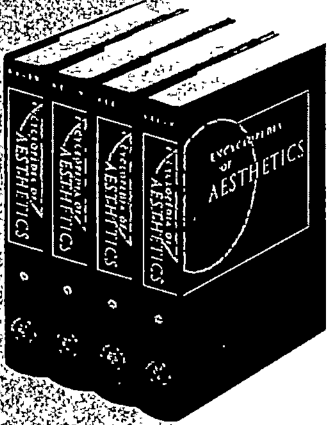
Prices are subject to change and apply only in the US.
To order, or for more information, please call
1-800-451-7556. In Canada, call 1-800-387-8020.
Visit our website at **www.oup.com**

OXFORD
UNIVERSITY PRESS

"An immediate reference classic."*

ENCYCLOPEDIA OF AESTHETICS

Editor in Chief: Michael Kelly



"*A unique work of splendid scholarship . . . highly recommended for any library serving a liberal arts program.*"—Choice

"*It is a magnificent achievement and deserves a very wide readership.*" —Journal of Philosophy

"*Both authoritative and accessible...an illuminating guide to historical and contemporary scholarship.*"—Library Journal

- 600 original articles
- 450 distinguished contributors
- 100 illustrations
- Bibliographies, cross-references, and an exhaustive index

In four volumes and 600 original articles, this landmark reference source surveys an extraordinary range of ideas about the arts in human life.

From Theodor Adorno to Frank Lloyd Wright, African aesthetics to video, the articles in the Encyclopedia examine a remarkable range of ideas and issues:

- the full breadth of critical thought on the arts—from classical philosophy to current literary theory
- in-depth survey of Western and non-Western traditions
- the political, social, economic, and ideological contexts of creativity
- the critical vocabularies of many forms, genres, and periods, from abstract expressionism to jazz to video
- the work of dozens of leading thinkers

2,240 pages • Four volumes
0-19-511307-1 • $495.00

OXFORD UNIVERSITY PRESS
198 Madison Avenue, New York, NY 10016
To order 1.800.451.7556 • www.oup-usa.org

Leigh S. Cauman

First-order Logic

An Introduction

1998 • approx. 336 pages • with numerous tables
Paperback approx. $29.95 • ISBN 3-11-015766-7

This textbook is designed to help its readers to reason systematically, reliably, and reflectively, in the course of their ordinary pursuits — primarily in inquiry and decision making. The principles and techniques recommended are explained and justified, not just stated; the aim is to teach orderly thinking, rather than the manipulation of symbols.

First-order Logic is presented in four parts, dealing with four major topics, in order of logical complexity (not in historical order): truth functions, monadic predicates, relations, and identity. Each part deals first with natural deduction (following Gentzen) and secondly with testing procedures, mainly truth trees (following Jeffrey). The reader is encouraged to see those two ways of looking at the material as complementary and to use them in tandem.

price subject to change

Walter de Gruyter
Berlin • New York

for North America:
Walter de Gruyter, Inc.
200 Saw Mill River Road
Hawthorne, NY 10532

Tel: (914) 747-0110 • Fax: (914) 747-1326

ANNOUNCING THE NEW

INTERNATIONAL ENCYCLOPEDIA OF THE Social & Behavioral Sciences

24 fully referenced Volumes plus 2 Index Volumes

A MILESTONE PUBLICATION - THE FIRST SOCIAL SCIENCES' MAJOR REFERENCE WORK FOR 30 YEARS

AVAILABLE IN PRINT AND ELECTRONICALLY!

The *International Encyclopedia of the Social & Behavioral Sciences* is the first attempt in a generation to map the social and behavioral sciences on a grand scale. Not since the publication in 1968 of the *International Encyclopedia of the Social Sciences*, edited by David L. Sills, has there been such an ambitious project to describe the state of the art in all the fields encompassed within the social and behavioral sciences.

The content of the Encyclopedia is being shaped to reflect:

- the astonishing growth and specialization of knowledge that has occurred within the disciplines since the 1960s
- the rapid development of interdisciplinary fields in tandem with the expansion of interest in policy and applications
- the internationalization of the research endeavor in response to the dynamics of globalism
- the impact of the computer and information revolutions on theory and practice within the social and behavioral sciences
- the new web of connections between the social and behavioral sciences on the one side and the biological life sciences on the other.

The Encyclopedia will be an essential resource for anyone studying and teaching within the social and behavioral sciences at the start of the new century.

Be sure to visit the Encyclopedia website regularly at:
http://www.elsevier.com/locate/iesbs
to keep up with developments

REQUEST FOR FURTHER INFORMATION

☐ Yes, please keep me updated on future developments with the *International Encyclopedia of the Social & Behavioral Sciences (IESBS)*

Name (please print) ____________
Position ____________
Department ____________
Organization ____________
Address ____________
Region/State ____________ Post/Zip code ____________
Country ____________
Telephone ____________ Fax ____________
E-mail/Internet ____________

Ref.0200029

ELSEVIER SCIENCE OFFICES

For customers in the Americas:

Elsevier Science
For the attention of:
Sally Iannacci
Marketing Department
P.O. Box 945
New York, NY 10159-0945, USA

For customers in other locations:

Elsevier Science
For the attention of:
Sally Iannacci
Marketing Department
P.O. Box 211
1000 AE Amsterdam
The Netherlands

Editors-in-Chief
(assisted by 50 section editors and an international advisory board)

Neil J. Smelser,
Center for Advanced Study in the Behavioral Sciences,
75 Alta Road, Stanford, CA 94305, USA

Paul B. Baltes,
Max Planck Institute for Human Development, Lentzeallee 94,
D-14195 Berlin, Germany

Biblio:
ISBN: 0-08-043076-7

Estimated publication date:
September 2001

Pre-Publication Price:
NLG 15,750.00 / US$7,995.00

Post-Publication Price:
NLG 19,690.00 / US$9,995.00

PERGAMON
An imprint of
Elsevier Science

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NUMBER 10
OCTOBER 2000

page
529 *Numbers as Types* J. R. Cameron
564 *Two Standpoints and the Belief in Freedom* Dana K. Nelkin

BOOK REVIEWS
577 Leigh S. Cauman: *First-order Logic: An Introduction*
Jonathan E. Adler

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President,* Arthur C. Danto; *Vice President,* Akeel Bilgrami; *Secretary,* Daniel Shapiro; *Treasurer,* Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc.

ISSN 0022–362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY
VOLUME XCVII, NO. 10, OCTOBER 2000

NUMBER AS TYPES

My thesis here is that talk about numbers is (albeit not recognized by us) a special kind of *type talk.* Numerical statements such as 'Three is an odd number', 'Five is greater than three', or 'Five plus three equals eight' are similar in kind to statements such as 'The whale is a marine mammal', 'The cheetah is faster than the tiger', or 'The engine and transmission together make up the power train'—where the instances of, say, the number five qua type (as all particular whales are the instances of the whale qua type) are all the particular collections made up of five things. In the material mode of speech, numbers are types, in a variant of the way in which the elephant, the electron, the double-reverse somersault, or the second Tuesday in the month is a type.

While the claim is initially about talk of numbers in everyday speech, it applies just as naturally to the most basic arithmetical statements, namely, singular statements about positive integers. It thereby answers a central question about this part of arithmetic, as to how it applies to experience, by tracing a direct link between such statements and statements about the collections of things of certain numerical sizes which we have to deal with in everyday life (five sheep, nine planets, three symphonies, and so on).

I shall later sketch how the analysis can be extended to quantified statements, and to talk about rational and real numbers, so as to explain how numerical mathematics generally, up to and including the level of real-number theory, can have application to our experience. And I shall also consider briefly how the analysis relates to the structuralist view of mathematics, and to some other accounts of what arithmetical statements are about. The first and main concern here, however, is to develop and defend the view that our ordinary talk about numbers, including arithmetical statements of the simplest

0022-362X/00/9710/529-63

kind, is a kind of type talk. That part of this article is an exercise in traditional nonformal conceptual or categorial analysis—conceptual and not just linguistic, concerned as much with how our thinking about numbers is articulated; though I shall for simplicity be focusing on our talk about numbers.[1]

Points regarding talk about types will sometimes be couched in the convenient shorthand of material-mode statements apparently about the types spoken of; this does not carry any commitment as to whether types in general, or numbers in particular, exist or do not exist as real entities of some kind.

We need first to examine type talk, and a few of the many forms it can take.

I. TYPE TALK AND DISTRIBUTIVE TYPE TALK

My concern is with just one way of using a descriptive referring expression as a type term, namely, using it with *distributive reference*, as in:

(A1) The blue-nosed whale eats 30 kilos of plankton per day.
(A2) Magnetic fields have an effect on the iron atom but not on the sulphur atom.
(A3) The Cross of Lorraine has a second small horizontal bar above the main one.
(A4) The blackbird sometimes exhibits albino coloring.

(Note the contrast between these and, for example, 'The dodo became extinct in the nineteenth century'—an example of use with *collective reference*—and more complex uses, as in 'Robert Thomson invented the pneumatic tyre', or 'The mobile phone is both a blessing and a curse'.) We use the device of saying 'The *F* is *G*', employing 'the *F*' as a type term distributively, as a way of saying of *F*s individually that each of them is *G*, or perhaps that every normal or typical *F* is *G* (the relevant restriction being understood from context); my main concern, however, will be with statements of unrestricted generality. While, as (A4) reminds us, universal application (unrestricted or restricted) is not invariably intended, it is the "default value," to be assumed in the absence of any explicit flag indicating otherwise. From here on, for brevity, I shall often use 'a type term' for 'a term used as a type term', and a restriction to use with distributive reference and intended universal application is to be understood.

A type term, 'the *F*', figuring as the only type term in a sentence, whether in subject position or elsewhere (as in (A2)), will be replace-able by 'any (normal/typical) *F*': it has widest scope. Relational statements involving two or more type terms—an important case for our concerns—are, we shall find, more complex to analyze.

---

[1] See further footnote 3 below.

While a type term in a language like English will most commonly take the form 'the *F*', involving the definite article, examples like these show the article is not essential:

(B1) Spring is a time of hope and expectation.
(B2) Man is capable of great tenderness but also of great cruelty.
(B3) Birth is an important event in small communities.

Other examples include talk about the days of the week or months of the year ('In January we ski'), quotation names of linguistic expressions, or the names of musical notes ('B flat', and the like), when used to refer to types rather than tokens. In languages with no definite article, type terms will of course regularly take this form. I shall be suggesting later that talk of numbers is a further example.

II. TYPES AND THEIR INSTANCES

How do we understand type terms used distributively? Consider first statements involving only one type term.

(i) *Types are instantiables, and universal.* 'The *F*', used as a type term, is treated as referring, or as if it referred, to something admitting of *instantiation* in particular *F*s, something on the "universal" side of the universal/particular divide, which has particular *F*s as instances or examples. If it is what we may call a *design type* (as in (A3) above), these instances will be seen also as being *tokens* of it.[2] The type is conceived not to have the kind of particular history or other individuating features that distinguish any one of its individual instances from every other instance, but to be capable of multiple realization in such particulars without being tied to any one of them. A design type (for example, the Irish Jig, the word 'nice', or the Paisley pattern) can have a history as a cultural object—even if it has no tokens, as in the case of a never-performed opera—and be a particular of its own special kind; but this a different way of being a particular from that of its tokens.

---

[2] The term 'design-type' is adapted from Rudolf Carnap's terminology in his *Introduction to Semantics* (Cambridge: Harvard, 1942), § 3. The type-token duality is a special case of the more general type-instance duality, distinguished by the type's being recognized as a pattern for agents to realize in particular performances or in objects they fashion, and for others in turn to recognize as realized in these tokens. Such a type is unusual in that (even though it is still essentially a pattern to be realized in tokens) it has a certain logical priority relative to its tokens, and an existence independent of their existence, as an object of thought (think for example of a never-performed opera, or a model of car designed but never put into production). The analysis of number talk to be offered below, it should be stressed, is in terms of the more general type-instance duality.

(ii) *Types are also quasi-particulars.* The *F* qua type differs from the abstract sortal universal *F*-ness in that, though not itself a particular with an individual history or set of particular features, it is treated as being capable of having some of the kinds of properties, including relational properties, that its instances can have. We say of the blue whale (though not of blue-whale-hood) that it swims, eats plankton, and suckles its young, just as particular blue whales do. Again, the isosceles triangle, we say, has three sides, has two of its angles equal, and is symmetric about one of its medians, just as any particular isosceles triangle is. A type is spoken of as if, somehow, it is both universal, not localized as its instances are, and yet also a kind of quasi-particular, sharing some of the characteristics these instances have.

(iii) *The type as embodied in its instances, or representative of them.* One way of understanding type talk is to think of a type as in some sense *embodied* in each of its instances (perhaps fully embodied only in the typical or normal ones), along with its characteristics (recall the etymology of 'instantiation' and 'instance'), so that whatever we attribute to it is thereby attributed to them. Another way is to think of it as a kind of *representative* of all its instances (recall some of the various senses of 'type'), so that, again, what we say about the type is thereby said of each of them. Whether either of these ways of thinking is to be taken seriously or is merely a *façon de penser,* they do contribute to shaping our understanding of talk and thinking framed in terms of types.

(iv) *An initial summary characterization.* Suppose, then, that '*F*' is a count noun (phrase), commonly used in expressions such as 'the *F*', 'this *F*', 'these *F*s', or the like in order to refer to some particular *F* or *F*s; then the expression 'the *F*', or '*F*' used on its own, is being employed in a given context not with particular reference but as a type term used distributively, with intended universal application, if there (I) we treat it as if it referred to something that both (a) admits of multiple instantiation, in particular *F*s, and hence has a different kind of identity from any individual *F*, and yet (b) at the same time is capable of having, and actually has, some of the same kinds of characteristics (including possibly relational ones) that its instances, particular *F*s, have; and (II) we understand what is said about it as being said in effect about each of these *F*s individually (or most of them, or typical ones, or...), perhaps on the basis of thinking of it as if it were embodied, with its characteristics, in all (or most, or typical, or...) individual *F*s, or as if it represented each of them individually.[3]

III. CATEGORY-STRADDLING TYPES

A given particular may be an instance of many different types: of the young male adult blue whale, the whale, the internationally protected marine creature, the mammal, and so on. Some of these types may fall into a hierarchy of types of progressively increasing abstraction, each a subtype of the type above it; a type high up in the hierarchy may have a very wide range of instances, with correspondingly fewer common characteristics shared by all of them.

An extreme case of this, that of *category-straddling types*, is of special interest here. While a type is conceived of as in some way akin to its concrete instances, and generally as sharing some of their characteristics, we sometimes recognize a type that abstracts from all the concrete properties of the things taken as its instances. We can do this where the instances, though they take quite diverse forms, perhaps even differing in category, still have something in common: perhaps some kind of similarity in structure, or a common function they serve in some use that we make of them. The obvious example is the case of words as types.

The word type 'giraffe' has as instances (in this case, tokens) both spoken utterances and inscriptions (the tokens of the spoken and the written word type, respectively). We find this quite unproblematic, despite the total lack of any physical characteristics shared by its two kinds of tokens, and indeed their being of fundamentally different categories (occurrents and continuants, respectively)—differences that mean that the word type itself is incapable of having either auditory or orthographic properties, and is indeed of indeterminate category (being neither an occurrent nor a continuant).

---

[3] Although the analysis leading up to this characterization may appear to be linguistic, and to focus heavily on English in particular, the criterion it offers is not linguistic but conceptual or logical, being framed in terms of understood content—effectively, in terms of the logical implications the statement in question is recognized to have. For example, it is by this criterion that we recognize that the expressions in the examples (B1) - (B3) in section I are functioning as type terms with distributive reference, despite the absence of the definite article.

At the same time, the account of number talk to be offered below is in general harmony on many points with the linguistics-based account of the uses of number words offered by James R. Hurford in his *Language and Number: The Emergence of a Cognitive System* (Malden, MA: Blackwell, 1987), chapters 4, 5. (He argues tellingly, regarding numerical adjectives (pp. 146-50 and 218-26), against assimilating them to quantifiers like 'all' and 'some', and also against the claim that they do not function as predicates—two suggestions I have not attempted to deal with here.)

The only kinds of characteristics which we can associate with 'giraffe' as a word type, those which the spoken and the written tokens can share, are structural and functional: an abstract structural property, reflecting the (supposed) correspondence between the temporal sound sequence in a spoken token and the spatial sequence of letters in a written token; and the functional properties (syntactic, semantic, and perhaps pragmatic) which the two kinds of tokens have in common as functionally equivalent components in the spoken and the written language, respectively, either capable of doing duty for the other (as when we write something instead of saying it, or vice versa). Despite the thinness of this set of characteristics, and the type's straddling two quite different categories in its tokens, we find no difficulty in thinking and speaking about 'giraffe' as a type, and in using this as a shorthand way of attributing characteristics simultaneously to its spoken and written tokens.

Numbers, I shall suggest presently, are another kind of category-straddling type.[4]

IV. MORE COMPLEX KINDS OF TYPE TALK

Finally, we need to look at more complex ways in which type terms are used distributively, including especially their use in relational statements.

(i) *Referring expressions formed out of type terms.* First, notice that, just as with any other referring expressions, we can form a new type term out of one or more other type terms, as in 'the nucleus of the hydrogen atom', 'the third Thursday of the month', 'the mating of the male horse and the female donkey', or 'the seven ages of Man' (this last serving to refer to a collection of types). These expressions have analogues, I shall suggest, in numerical expressions such as 'the square of seven', 'the sum of three and five', or 'the prime factors of eighteen'.

(ii) *Statements involving two or more type terms.* A statement may involve two or more type terms, all used distributively, as in:

(C1) The Montague hates the Capulet.
(C2) The puma is between the civet and the leopard in size.
(C3) The legal guardian is responsible for the ward's upbringing.
(C4) The written examination, the submission and assessment of a thesis, and the *viva voce* examination together constitute the final examination for the degree.

---

[4] I would also want to argue that talk about propositions has evolved out of talk about another kind of category-straddling type, one that has as its instances both assertions that people make and beliefs that people hold.

Type terms in such statements may function in different ways. 'The *F*' in subject position in such a statement will (unless otherwise indicated) serve to say something about all (normal/typical) *F*s; but in other grammatical positions there are other possibilities. My interest is in two cases.

(a) *Multiply universal type talk.* What is said in (C1), for example, applies to any (true) Montague and any (true) Capulet in relation to each other: both type terms are used distributively with (possibly restricted) universal application. The three type terms in (C2) function similarly.

(b) *Matching-pair and matching-n-tuple universal type talk.* (C3) asserts that any guardian is responsible for the upbringing, not of any and every child, but of the corresponding child or children. (The kind of correspondence involved in cases of this kind may be one-one, one-many, many-one, or many-many.) Here, 'The *F* φs the *G*' is understood, on the basis of a correspondence between particular *F*s and *G*s which is usually to be recognized from context, as meaning 'The *F* φs the corresponding *G*', this being taken as equivalent to 'Each *F* φs each of the *G*s corresponding to it'. We can think of such a statement as equivalent to a universal statement about a *matching-pair type,* the-*F*-and-corresponding-*G*, to the effect that its *F*-member φs its *G*-member, any particular pair $\langle f, g\rangle$ being an instance of this pair type if *f* is an *F* and *g* is a *G* corresponding to that *F*. (C4) exemplifies a like pattern involving more than two type terms, and can be seen as a statement about a matching-quadruple type.

We can bring cases (a) and (b) together under a single characterization if we think of any instance of case (a), a simple multiply universal statement, as involving a matching-pair type or matching-*n*-tuple type for which the correspondence requirement is vacuous—is satisfied by any *F* and any *G* whatever (or any *F*, *G*, *H*,...whatever). And we can construe a statement of either kind, 'The *F* φs the *G*', taken as being about a pair type, the-*F*-and-(corresponding)-*G*, in terms of either the "embodiment" conception or the "representative" conception mentioned earlier (section II, (iii)). The characterization of the use of terms as type terms with universal distributive reference offered in (iv) of section II can to be modified to accommodate this kind of statement.[5]

---

[5] Modified, it might run like this: where one or more count noun (phrases), '*F*', '*G*', '*H*',..., in the singular (and in English, with or without the definite article prefixed), are used together in a single statement, they are each of them being employed there as type terms with distributive reference if there: (I) we treat each as if it referred to something which both (a)... and (b)... [features (a) and (b) being

(iii) *Type terms figuring in identity statements.* To round off, we should notice the use of two type terms with distributive reference in generic identity statements, such as:

(D1) The mule is the issue of the mating of the male donkey with the female horse.
(D2) The rainbow is the visual effect produced when the spectrum of different wavelengths of light in sunlight is differentially refracted within falling raindrops.

(C4) above can also be taken as asserting a type-level identity, between the members of a matching-triple type, taken together (the written examination, the thesis, and the *viva voce* examination, in combination), and a type (the final examination for the degree). In a statement of this kind, we assert the numerical identity of two types, 'The *F* is the *G*', as a way of affirming that a generic identity (whether analytic, or synthetic though holding universally, as in (D2)) holds at the level of the corresponding particulars: every *F* is a *G*, and conversely. Usually the context will indicate how, for a given *F* (or *G*), the relevant *G* (or *F*) is to be identified. Such statements admit readily of interpretation in terms of the "embodiment" way of understanding type talk, but perhaps not so straightforwardly in terms of the "representative" interpretation.

V. NUMBERS AS TYPES: WHAT THE THESIS IS

By now it will be obvious in a general way what is meant by the claim that talk about numbers is a kind of type talk.[6] The claim, initially, is

---

as in (I) of the original characterization]; and (II) we understand what is said about the ordered *n*-tuple of the things to which the terms are respectively taken to refer as effectively being said about the members of all (or most, or...) corresponding particular matched *F-G-H-...n*-tuples (the matching required being understood from context or in some other way), perhaps treating the *n*-tuple type as if it was embodied in each of these particular *n*-tuples, or as if it represented each of them individually.

[6] As already indicated (footnote 2), the claim is that numbers are types of the kind that have instances, rather than design types whose instances are treated as tokens of themselves.

Charles Parsons has at various times discussed a view of numbers as types that have, for example, strings of corresponding numbers of strokes as tokens—this in the context of exploring a conception of intuitive awareness of mathematical objects—"Mathematical Intuition," *Proceedings of the Aristotelian Society*, LXXX (1979-80): 145-68, and "Intuition in Constructive Mathematics," in Jeremy Butterfield, ed., *Language, Mind and Logic* (New York: Cambridge, 1986), pp. 211-29. But his concern is not, as ours is here, with the analysis of *talk about* numbers as a kind of type talk. And his reasons for ultimate dissatisfaction with the view of numbers as types do not bear on the claim being made here, I believe. In whatever way he may be using the term 'token', some of these reasons apply only on an interpretation of numbers as design types; and others relate to his interest in the types in question as about our everyday use of number words as nouns (rather than as adjectives), a use corresponding to the statements of what we might call *basic arithmetic*, that is, arithmetic confined to the positive integers; the term 'number word' will be used with this sense from here on, until late in section VII. The claim is that in saying, for example, 'Five is an even number', 'Five is greater than three', or 'Three plus five equals eight', we are using the number words as type terms with distributive reference, in order to say something universal about particular collections of the corresponding numerical sizes—such things as five sheep, or three somersaults, or eight operas. Let us for convenience call such collections "fivesomes," "threesomes," and so on, without inquiring at this stage what kind of things these are.

In every respect except one, to be noted in section VII (1) and discussed in section VIII, the use of the words 'seven', 'five', and 'twelve' in the statement 'Seven plus five equals twelve', for example, clearly fits the characterization of type terms used distributively which was offered earlier (perhaps in the extended version envisaged in section IV (ii)). (I shall examine this kind of statement further in section VII.) 'Seven' in this kind of use is taken to denote something which, on the one hand, is not a particular, but is instantiated in particulars—in the buttons on my shirt, the days of the week, the Wonders of the Ancient World, the Deadly Sins, and so on—and yet, on the other hand, is also capable of having some of the characteristics that these particulars can have—being odd or even, for example. And attributing any particular characteristic to seven will in effect be attributing the same characteristic, or a corresponding one, to all sevensomes; for example, to say that seven is odd is to say that any sevensome of *F*s (whatever kind of things *F*s may be) is odd—that when we pair off the *F*s, there will be one left over.

In the case of many characteristics, a special mathematical term—for example, 'prime'—has evolved for the characteristic as we at-

---

possible objects of a kind of intuition that would yield mathematical knowledge. (The account to be offered below suggests an alternative source of knowledge of the truths of numerical mathematics.)

David Bostock considers briefly (and dismisses) what is in effect, though not expressed as such, the suggestion that 'two' might function as a type term—*Logic and Arithmetic* (New York: Oxford, 1974), pp. 3-4. But it is the use with collective reference, to refer to the class of all pairs of things, which is in question there, not the use with distributive reference, as here. (What Bostock's own analysis of numerals as a form of quantifiers in fact analyzes is the everyday use of numerals *as adjectives*, that is, their use coupled with suitable sortal noun expressions, to refer to what will be identified shortly in the text as pluralities; his account is entirely compatible with the view that number nouns and numerals function in arithmetical statements as type terms.)

tribute it specifically to numbers (the types), and we may hesitate to apply the very same term to the corresponding particular collections; but an exactly corresponding characteristic will be attributable, under a different label. (In linguistics, similarly, many terms, such as 'polysyllabic' or 'loan word', have been coined for application exclusively to word types, though what we say in using them relates ultimately to tokens of these words.)

The account of number talk as a form of type talk—the *type analysis*, to give it a label—surely has a strong initial plausibility. One immediate major point in its favor noted already is that it offers a very simple and natural answer to a fundamental question about basic arithmetic: How do its statements about numbers apply to the numerical quantities we encounter in everyday experience? (Or, equivalently, the question: How is our use of number nouns linked with our use of numerical adjectives?) The account needs now to be fleshed out, however, first by identifying what kind thing a number qua type has as its instances—what threesomes of Fs, fivesomes of Gs, and so on are—and then by examining more closely just how number talk functions: how we are able to talk about such collections, and relations between them, by making statements about numbers. We shall also want to understand how we can know the truth or falsity of statements of this kind. A basic challenge to the thesis will then need to be considered in section VIII.

VI. COLLECTIONS AND COUNTING; PLURALITIES; NUMBERS AS PLURALITY TYPES

A good way to determine what kind of things threesomes, fivesomes, and so on are is to look at the procedure of counting, since counting is the basis for identifying the numerical size of any but the very smallest collection. We need in any case to look at counting, in order to appreciate how arithmetical truths can have the special kind of certainty they exhibit, and how we are able to grasp this certainty.

When we count, we are determining how many Fs there are in some aggregate or totality, where Fs are some kind of countable thing.[7] A central feature of counting is that, as Gottlob Frege[8] noted, it involves invoking some principle of individuation and counting, and the numerical attributions to which it leads are not absolute but, as we might say, *sortal relative.* A question asked about a given aggregate—'How many are there?'—will have a clear meaning, and an answer, only as the question 'How many *F*s?', where *F*s are some kind of thing for which there is a principle for individuating and counting; and the answer may vary for different *F*s: the troops on parade may be one regiment, two battalions, four companies, or two hundred soldiers.

---

[7] 'Aggregate' for this purpose is to be understood very broadly, and certainly as including, for example, aggregates of scattered individuals; its range should accord with the very wide range of kinds of things we can count.

[8] *The Foundations of Arithmetic*, J. L. Austin, trans. (Malden, MA: Blackwell, 1953), § 22. The point (attributed by Frege to J. J. Baumann) was anticipated by Plato in the *Meno*, Aristotle in the *Metaphysics*, and George Berkeley in the *New Theory of Vision*, and is closely linked with John Locke's thesis that identity is sortal relative (*An Essay Concerning Human Understanding*, book II, chapter 27, § 7). Use of the term 'sortal' here is convenient but possibly not entirely accurate: principles for

Counting thus involves applying to some aggregate or totality a principle for individuating and counting some kind of countable things, *F*s. Consider, first, the simple kind of case, where the principle in question requires that individuals, to be distinct, must be mutually disjoint in their composition. Here, if the aggregate admits of being seen (among the different possible ways of decomposing it in thought) as composed precisely and without remainder, and in a unique way,[9] of a number of *F*s, then when we do think of it specifically as composed of these *F*s, we are treating it as what we may call a *plurality* of *F*s. If we regard a woodpile in a sawmill as consisting of a number of logs (rather than as, say, a quantity of timber, or as made up of so much cellulose, water, and so on), we are choosing specifically to see it as a plurality of logs, taking the aggregate, we might say, *sub specie principii individuationis.* 'Plurality', as used here, may be defined, in this simple kind of case, as *an aggregate taken relative to some appropriate principle for individuating and counting,* where an appropriate principle will be one relative to which the aggregate can be seen in the way indicated above.[10] (More on pluralities shortly.)

---

counting come in very varied kinds, some of which may not respect the boundaries between categories or sorts.

[9] There are conditions for *F*s to be countable in general, and conditions for a given aggregate to be seen as a totality of *F*s; the uniqueness requirement may involve both. (That requirement will fail, for example, where a journey can be seen as made up of stages in various different ways, involving different numbers of stages.) Another important part of the uniqueness requirement is that the result of counting should be *order indifferent,* being the same regardless of the order in which the *F*s are counted (this might fail to hold, for example, where we are counting cells which are dividing rapidly). A much more systematic exploration than is possible here of these conditions, and of the counting procedure itself, is obviously required in order to establish fully the claim made below that our understanding of them is the basis for our grasp of arithmetical truths.

[10] The distinction made here between aggregates and pluralities is close to Hurford's aggregate/collection distinction (*op. cit.,* pp. 137-38), though not defined in the same way.

The actual counting will consist in systematically identifying the individual *F*s one by one, pairing them off as we go along with *tallies* of some kind, which may be numerals or number words, but need not be. When we have worked through to the last *F* in the plurality, some rule will tell us, from the position reached in respect of the tallies paired with the *F*s, what the number of *F*s is. For example, the tallies may be strokes, one for each *F*, which we have grouped in tens as we went along; if by a further, shorter count we find there are six such tens, with four strokes left over, we know there are sixty-four *F*s in the aggregate. If instead we have paired successively identified *F*s with the number words 'one', 'two', and so on, in order, then if the last *F* was paired with, say, 'seventeen', we know there are seventeen *F*s. And what we are attributing the numerical property to is the plurality, the totality of *F*s *taken as* a totality made up of *F*s.

Note that—contrary to Bertrand Russell's[11] view—there is no reason to regard counting as correlating objects with numbers. The natural account of the use of the word 'eleven', say, in counting is that it is functioning within a procedure which may happen to employ words or numerals but need not, and, where it does use the word 'eleven', uses it as a tally and not as a noun denoting a number, nor as any other part of speech for that matter—since it is not using it within a sentence. (If anything, its function in this use is linked rather with the corresponding cardinal or ordinal adjective than with the noun, since what its use in effect registers is that eleven *F*s have so far been counted, or that the eleventh *F* is being counted.)

The concept of a plurality which is in play here is not as simple as it may appear; but we need to grasp that pluralities are for us the most ordinary of things, part of the stuff of our everyday experience as we speak and think about it, in that they are what the plural referring expressions that we use constantly ('Tom, Dick, and Harriet', 'the outer planets', 'these five cows') serve to pick out. In the simple case considered so far, where distinctness entails mutual disjointness, a plurality is identical with an aggregate regarded in a certain way, something relatively concrete, not more abstract than the things of which it is composed. (While a conceptual input from ourselves is involved in our applying a principle of individuation and counting, in doing this we are merely focusing on one of the many aspects that the aggregate does itself actually possess. And we do the very same sort of thing in picking out a single individual in order to refer to it.) To reflect the special recognition we accord to the *F*s making up an aggregate when we take that aggregate as a plurality of *F*s, we may call them the plurality's *constituents*.

---

[11] *The Principles of Mathematics* (London: Allen and Unwin, 1937), § 129, where he was arguing that the attempt to give an account of numbers by reference to counting proceeds in the wrong direction.

Matters are more complex in the case where the principle for individuating *F*s does not require mutual disjointness as essential for mutual distinctness: think, for example, of how we individuate the words in a crossword, or the rooms in a building (when each room is taken as including the possibly shared enclosing walls, ceiling, and floor), or how we count the (eight) triangular areas in a diagram consisting of a square with the diagonals drawn in. Here we cannot regard the plurality of the *F*s in question as the *initial* aggregate—the aggregate to which the principle for individuating *F*s was originally applied—taken relative to that principle (since two or more clearly different pluralities of *F*s could each be identical with the same initial aggregate taken relative to that principle). Rather, we must see it as the *subsequent* totality of the individual *F*s, identified *after* these have been singled out individually.[12] But what matters for our purposes is that in this kind of case, just as much as in the simple case, the plurality of *F*s is the same kind of thing as a single *F* except in being several where the latter is only one. Just as 'these five cows' denotes more of the same as what 'that cow' denotes, so 'these five rooms' denotes more of the same kind of thing that is denoted by 'this room', and not something of a quite different and more abstract nature.[13]

---

[12] The *F*s of which it is composed are then treated effectively as if they were *atomic* parts of this subsequent totality: a proper part of an *F* is not treated as a part of the totality, unless it itself happens also to be an *F* (whereas it will always be treated as a part of the initial aggregate). What allows us to differentiate in this way between an *F* and its parts in relation to the plurality is the fact that the individual identity of each *F* is grounded in some kind of unity which it, the whole, has but its parts in general do not have (for example, the unity of a complete word, a room, or a triangular area); this yields one sense of 'part' in which the *F*s, the constituents of the plurality, are parts of the plurality whereas (in general) their proper parts are not.

Further on pluralities of this more complex kind, see the discussion in my "Plural Reference," *Ratio* (new series), XII, 2 (June 1999): 128-47, note 5 (pp. 135-36), though I am no longer wholly confident about what is said there. It is because of this kind of case that, when considering counting and numerical attributions, the role a principle of counting has in the identifying of any plurality needs to be stressed strongly—more strongly than is usually done in discussions of pluralities, and even in Peter Simon's discussion in his otherwise very full treatment of every kind of mereological concept in *Parts: A Study in Ontology* (New York: Oxford, 1987), §§ 4.3-4.4.

[13] We do not need here to try to decide whether a plurality of *F*s is something different from the set having these *F*s as its members. But whatever we decide about sets, it remains clear that an instance of five, such as a fivesome of cows, is just these cows as a totality (taken specifically as a totality *of cows*), something relatively

It is to pluralities, so understood, that we attribute numerical sizes: threesomes, fivesomes, and so on are pluralities. And the claim made here is that talk about the number five, for example, is talk about a plurality type, the fivesome, whose instances are all particular fivesomes, all those pluralities which have five constituents. More precisely, while there are different kinds of plurality types, our concern is with *size-differentiated* plurality types (in contrast, for example, to types of plurality distinguished by the kind of things of which their instances are composed, such as the-plurality-of-dogs, of which any particular plurality of dogs is an instance). The contention is, then, that talk about numbers is talk about such size-differentiated plurality types, or (in the material mode) that numbers are plurality types of this kind—the threesome as type, the fivesome as type, and so on. The term 'plurality type' will be used in this sense of 'size-differentiated plurality type' from now on.

VII. NUMBER TALK AS A PARTICULAR FORM OF TYPE TALK

How, more precisely and in terms of the account of type talk offered earlier, do statements about numbers say something about particular pluralities by saying something about the corresponding plurality types? Before analyzing an example, let us note some general features of number talk as type talk, starting with the one problematic feature alluded to earlier, and two unusual but unproblematic ones.

(1) A number noun such as 'five' diverges from the standard pattern associated with type terms in not having a systematic duality of function: it does not admit of being used just as readily either with particular reference or with type reference. We can and sometimes do employ expressions like 'a five', 'this five', and so on to refer to a particular fivesome (see, for example, the talk about 'tens' in the fourth paragraph of section VI above; we also very commonly talk in this way when teaching children arithmetic). But we use number nouns in this way rather infrequently and not routinely. The relative rarity of this use of number words as nouns with particular reference, I shall suggest, is what has prevented number talk being recognized as a kind of type talk; we shall need in due course to find an explanation for it.

(2) Recall next that while type terms in languages like English usually include the definite article, they need not do so. And much as the terms 'Spring', 'Man', or 'birth' function as type terms in the

---

concrete and not an entity of some quite different kind. (See "Plural Reference," especially §§ 2-5, for discussion of the question whether pluralities and sets are different kinds of things.)

statements in (B1) - (B3) above, the terms 'three', 'five', and 'eight' function as type terms in 'Three plus five equals eight', despite their not having what is in English the more standard form, 'the *F*'.

(3) Another distinctive and very important feature of talk about numbers is that the types it speaks of, like words, are category-straddling types—types, however, that straddle not just two categories, but all categories of countable things. The instances of seven will be sevensomes made up of constituents of any (countable) kind whatever. We saw that what underlies and gives point to our saying things in a compendious way about both spoken and written tokens of 'giraffe' by means of statements about the word type is the common structure and function shared by all instances of the type, whether spoken or written; what is the corresponding feature in the numerical case, in virtue of which we can similarly frame universal statements about all sevensomes, of any and every kind, in terms of seven or the sevensome? To say that it is their common cardinality leaves all the explaining still to be done. How can things of such disparate kinds have, and be recognized to have, something in common? And why does this shared feature merit recognition?

The answers lie in the uniformity of the counting procedure, whether it is being applied to frogs or fjords, financiers or fanfares—in the fact that the procedure is exactly the same no matter what kind of things are being counted, with its corollary that whatever follows about one totality of *F*s from discovering that it is (say) a sevensome will follow in like manner about any other plurality of *F*s from the same discovery about it, and equally about any plurality of *G*s from the same discovery about that plurality, no matter how categorially disparate *F*s and *G*s may be. And there are many things that follow from something's being a sevensome: for example, that it is an odd plurality (in the numerical sense); that it is a prime plurality (that is, does not admit of division without remainder into equal parts containing more than one constituent each); and that it will stand in various relations to pluralities of other sizes. So it is natural for us to pay attention to the features which go with being a sevensome and which will be possessed by a sevensome of any kind whatever.

(4) This is why there are true universal statements about all sevensomes, of any and every kind; and we make such statements by making the corresponding statement about the sevensome as a type, that is, about seven. The usefulness of such statements, not least in teaching, makes this use of number nouns as type terms very valuable, and one which we constantly exploit.

(5) Further, we can begin to see already that what enables us to grasp these universal truths is just our understanding of what is

involved in counting (including the conditions of countability), and of the implications of numerical attributions based on counting; this will become clearer as we proceed. It is in this direction that we should look for an explanation of how we can have a priori knowledge of such truths.

(6) The truths of basic arithmetic, framed as statements about plurality types, are really about the instances of these types, particular pluralities. But because plurality types straddle the categories of every kind of countable thing, these truths are not about particular pluralities of any one kind, and their particular properties and relations, but equally about pluralities of any and every kind, collections made up of any kind of countable things whatever. They are in fact about a pattern of properties and relations that go with being pluralities of the relevant numerical sizes, irrespective of the kind of countable constituents of which these pluralities are made up—what we might naturally call a *structural* pattern of properties and relations associated with numerical size.[14] This finding fits in very naturally with the structuralist view of the nature of mathematics generally, the view (putting it very summarily) that a mathematical theory is not really about objects of a particular kind, and the particular properties and relations these objects exhibit, but is in effect about a pattern of relationships holding within different possible systems of objects, systems which may be made up of very different kinds of objects, but are all isomorphic in respect of that pattern.[15] (More on structuralism and the type analysis in section XI (ii) below.)

(7) How does this account apply to particular statements about numbers? Consider, for example: (E) Three plus five equals eight. We have seen that (E) might be expressed as 'The threesome plus the fivesome equals the eightsome'. What does this assert about particular pluralities, and how does it do this? Two analyses are possible, the simpler taking 'equals' to express equality in numerical size or equicardinality, the other taking it to express identity.

---

[14] In fact, what seem to be statements about properties of numbers such as being even, or being prime, prove on probing to be implicitly relational in nature. This yields another sense in which statements about numbers are structural, one which needs to be distinguished from the wider sense linked with the straddling of categories.

[15] See, for example, Stewart Shapiro, "Mathematics and Reality," *Philosophy of Science*, L, 4 (December 1983): 523-48, especially section III, and other references given there. Different versions of this view, and their problems, are discussed by Parsons in "The Structuralist View of Mathematical Objects," *Synthese*, LXXXIV, 3 (September 1990): 303-44.

Two preliminaries: first, note that we can understand the use of functional expressions such as 'the square of...' or 'the sum of...and...' with number nouns, to create terms that may then themselves be used as type terms, as analogous to the use of type terms formed out of other type terms which was noted in section IV (i). Also, let us describe two pluralities as "co-countable" if they are made up of the same kind of constituents, with the same principle of counting in play in respect of each of them.

(i) On the "equality in number" interpretation, and taking 'three plus five' as a functional expression, (E) attributes to one plurality type, the union of the co-countable and mutually disjoint threesome and fivesome, numerical equality with another plurality type, the eightsome (where the requirements of co-countability and mutual disjointness are built into the meaning of the functional expression 'plus'). And what it thereby says about particular pluralities is that any plurality that is a union of a particular co-countable and mutually disjoint threesome and fivesome is equal in size to any particular eightsome whatever. This analysis treats the arithmetical statement as a straightforward multiply general type statement (an example of case (a) in section IV (ii) above), like example (C1) but having a functional expression as one of the type terms.

Alternatively, we may take (E) not as involving a functional expression, but as being about a single matching-triple type rather than two types (exemplifying case (b) in section IV (ii)), asserting of the appropriately matched twosome, threesome, and fivesome, that the first two, taken together, will be equal in numerical size to the third. The matching required in any instance of the triple type will be that the first two be co-countable and disjoint, with a vacuous requirement regarding the third. (Again, these requirements of matching are built into the term 'plus'.)

The analyses are obviously equivalent, and, on either version, we get just what we would expect as the account of what (E) says about particular countable collections. (Note that on either version, there is no implication that the eightsome must be co-countable with the other plurality: three bachelors together with five other bachelors are equal in number to eight weddings, for example).

(ii) On the identity interpretation of 'equals' (reflected in the colloquial 'Three plus five is (are) eight'), (E) presents a certain complex type, the union of the co-countable and mutually disjoint threesome and fivesome, as identical with another type, the eightsome. Asserting numerical identity of two types, as we saw in section IV (iii), is asserting a generic identity at the level of instances; in this case, it is asserting that every union of a particular co-countable and mutually disjoint threesome and fivesome is a (particular) eightsome. As in the equality interpretation, an alternative version of this interpretation is possible (corresponding to the first reading of (C4) in section IV (ii) (b)), which would treat the statement as being about a matching-triple type, the suitably matching threesome, fivesome, and eightsome (with rather more complex matching requirements for instances of this triple), and asserting that taking the threesome and fivesome together as a single plurality yields the eightsome.

Either version again yields just the account of what (E) asserts about particular pluralities that we would expect. The apparent difference between an interpretation in terms of equality in size and one in terms of identity is purely nominal: the various analyses of what is said about particular pluralities are all logically equivalent (ultimately because a particular plurality's being equal in size to any and every eightsome is in fact the same thing as its being itself an eightsome). Of the two analyses, however, an equality analysis provides a better basis for understanding statements of comparative magnitude, such as 'Three plus eight is greater than nine'.[16]

As I suggested already, we recognize the truth of a statement such as (E) just from our understanding of the counting procedure, by reflecting on what the result must be when we count a plurality that is the union of a mutually disjoint threesome and a fivesome: when to a threesome we add a onesome, the result must be a foursome; so when we add instead a twosome, the result can only be a fivesome; and so on, until we see that adding a fivesome must yield an eightsome. Further, we can then go on, through a reflection of a more general kind, to recognize systematic truths: for example, we can grasp that addition is always commutative ($m + n = n + m$, for any m and n), simply from understanding that the result of any counting whatever must be indifferent to the order in which we count the constituents of the plurality—this order indifference being a necessary condition of countability. (The associativity of addition can be recognized similarly.)

(8) While it is fairly easy to see in general how the analysis will extend to other kinds of statements in basic arithmetic, the case of multiplication statements warrants a brief look. The most natural way of understanding what 'Three times five equals fifteen' says about

---

[16] Each analysis has its place: the equality analysis fits, for example, the case where we are checking that we have enough lunch packs for everyone in the class to have one each, the identity analysis the case where we are keeping track of eight children being transported in two cars, three in one and five in the other, and we need to be sure we have, not just any eight children, but the right eight.

pluralities of *F*s takes it, not as being about the result of "multiplying" a threesome of *F*s by a fivesome of *F*s, but rather as being about the total number of *F*s there are in a plurality made up of three distinct fivesomes of *F*s: a threesome of fivesomes of *F*s will, taken as a single plurality of *F*s, be equal in size to any fifteensome (on an equality analysis), being itself a fifteensome of *F*s (on an identity analysis). (Compare the colloquial paraphrase 'Three fives are fifteen'.) On this analysis, which acknowledges the familiar point that multiplication is in fact a kind of repeated addition, the two multiplicands have different roles—a fact easily overlooked once we become familiar with the commutativity of multiplication. Commutativity here has a rather different basis from the commutativity of addition; but like the latter it can be deduced from the nature of counting—in particular, from the fact that the result of counting must be order indifferent, and will be the same no matter what kind of thing is being counted.[17]

(9) The type analysis can readily accommodate zero and the negative integers, and so extend from basic arithmetic to embrace the arithmetic of the integers generally. It is not hard to see how we may come to think in terms of a null plurality of *F*s, or a negative plurality of *F*s—for example, when we are making inventories, comparing numbers of things, setting surpluses against shortfalls or assets against

---

[17] The simplest way for us to grasp that, for example, three times five will always be the same as five times three is from a recognition that some of the relevant instantiations will be objects admitting of being arranged in a three-by-five array: knowing that the result of counting such an array must be order indifferent, we can appreciate that the result of counting will be the same whether we treat it as a threesome of fivesomes, counting through the fivesomes in turn, or as a fivesome of threesomes, counting through the threesomes in turn. And having grasped that commutativity holds in these cases, we can appreciate that it will hold in all cases of three times five, because the outcome of a counting procedure will be the same no matter what kind of things are being counted. Then, having grasped that $m \times n = n \times m$ for some particular m and n, we can appreciate that like reasoning will establish the same for any m and n. By analogous reasoning, we can recognize the associativity of multiplication, and the distribution law $(m \times (n + p) = (m \times n) + (m \times p))$.

Notice that the account of multiplication offered here—like that provided by John Bigelow in *The Reality of Numbers* (New York: Oxford, 1988), pp. 83-85—is wholly general, and not tied to a "geometrical" interpretation, which treats the product as made up of a different kind of thing from the things multiplied together (as, for example, units of area differ from units of length), though it does also accommodate this case. There are other possible ways of defining multiplication formally within arithmetic, on some of which (including Bigelow's) its commutativity is trivially obvious; but they will not yield as simple an account of how '3 × 5 = 15', for example, tells us something about particular pluralities, or provide as natural a basis for understanding in turn the arithmetical operation of division, and how it relates to possible experience.

liabilities, and so on; then we shall come to talk about zero and negative numbers as the corresponding plurality types.[18]

(10) We can understand arithmetical statements, as we can other kinds of type talk, in terms of the "embodiment" conception of types, for example, taking (E) as saying that wherever a certain matching-pair type, the-mutually-disjoint-threesome-and-fivesome, is embodied in a pair of pluralities (of *F*s), another type, the eightsome, will also be embodied in the union of the pair (qua plurality of *F*s). We could equally apply the "representative" interpretation of type talk to most statements about numbers (though perhaps not so simply to such statements taken as identity statements: recall the remark at the end of section IV).

(11) Finally, notice that while distributive talk about 'the *F*' as a type is often indeterminate as regards the range of particular *F*s to which it applies (the "true" or "proper" ones), true statements about numbers as types, as here analyzed, apply with strict universality, regardless of context. Failure of some arithmetical statement to hold good of some particular *n*-some of *F*s will be taken to indicate that these *F*s, or perhaps *F*s generally, do not in fact satisfy the conditions for countability. Again, the requirements of matching, for example, for (E) to apply to a given threesome, fivesome, and eightsome are quite clear and determinate, independently of context, being implicit in the meanings of 'plus' and 'equals'.

And as I have been suggesting, what gives everyday arithmetical statements this strict universal validity is their following from the nature of the counting procedure (including the conditions for its applicability), while their peculiar epistemic certainty is grounded in our ability to deduce their truth from our grasp of that procedure and what it entails about the results of particular acts of counting, when applied to collections that are genuinely countable. (What makes a statement about a particular (supposed) plurality contingent is that it is a contingent matter whether the plurality's constituents really are countable; this also, together with our fallibility in counting and calculating, is what renders our making of such statements fallible.)

---

[18] Frege held that there could not be such a thing as a null aggregate: "if we burn down all the trees of a wood, we thereby burn down the wood"—Peter Geach and Max Black, trans. and eds., *Translations from the Philosophical Writings of Gottlob Frege* (Malden, MA: Blackwell, 1952), p. 89. Whether or not he was right about null aggregates, we can and do talk quite naturally in terms of null *pluralities* (without necessarily implying that they really exist): for example, in making an inventory of the number of trees in different plantations, we might record "none" in respect of one where they have all been burned down.

VIII. A PROBLEM FOR THE CLAIM

One difficulty faces this analysis of number talk. Where a term is being used as a type term, this will have to be evident to users of the language generally if the use is to be recognized for what it is and understood correctly. So if number talk really is a kind of type talk, should that not have been obvious, and any argument for the claim quite unnecessary? (Does the need to argue for the claim not show that it must be incorrect?) If talk about numbers is a form of type talk, a satisfactory explanation is needed of how we could fail to realize this fact. The required explanation unfolds in stages.

(1) It seems clear that what has hindered recognition of number talk as talk framed in terms of types is the anomalous feature alluded to already, the absence of a routine type-instance duality in the use of number words as nouns. Whereas standardly our use of a term of the form 'the *F*' or '*F*' as a type term accompanies and complements a contrasting use of 'an *F*', 'these *F*s', and so on as expressions with particular reference, we shall only occasionally refer to a particular fivesome as 'that five', and probably with some sense of unease when we do. The type-instance duality and the contrast between the two ways of using the same term are integral to our recognition and understanding of type talk in all other cases (as the characterization in (iv) of section III acknowledged); its relative rarity in this case is what has prevented talk about numbers being recognized as a variety of type talk.

(2) To convert this diagnosis into an explanation, however, we need to identify the source of our hesitancy about using number nouns with particular reference. That source is readily located: we feel uneasy in referring to a particular fivesome of *F*s as 'this five', using 'five' as a noun, because we are torn between thinking that 'five' here should be a singular noun and thinking that it should be plural—torn between saying 'this five' and saying 'these five'. Given this uncertainty, our unease about using 'five' as a singular noun referring to a particular threesome is hardly surprising; rather, it is more surprising that despite the uncertainty, we do still sometimes use it in this way.

(3) This explanation will be more convincing if we can identify what gives rise to our uncertainty here. The source lies in the inflexible way in which the grammatical singular/plural distinction applies to expressions referring to pluralities, which fails to match the more sophisticated and flexible way in which we actually think about pluralities.

In referring to a plurality ('the Smith brothers', say) we can recognize it, with no inconsistency, either as many things, in a reference that is to be understood distributively (as in 'The Smith brothers are fair-haired'), or as one thing, in a reference to be understood collectively ('The Smith brothers run a large estate'); sometimes we have to regard it simultaneously as both many in one perspective and one in another, in a reference that has both a distributive and a collective element ('The Smith brothers work as a team').[19] This numerical duality should not be found surprising, given that numerical attributions are always relative to the criterion of counting we are employing. And, in fact, both aspects of the duality are involved when we attribute a numerical size to a plurality, since, on the one hand, it is the *F*s together, the (one) plurality, which is nineteen or ninety-eight in number, but, at the same time, it has this cardinality in virtue of its multiplicity—its being the many *F*s in question: referring to an *n*-some necessarily involves combining the thought of it as one thing with a thought of it as also being (relative to its associated principle of individuation) many.

Grammar, in contrast, is (in many languages at least) inflexible about oneness and multiplicity, requiring us when we speak about a particular plurality to make a choice between treating it exclusively as one thing (using the singular) and treating it exclusively as many (using the plural). In particular, if in English we want to use 'five' as a count noun with particular reference (for example, referring to 'the nearest five', intending 'five' to be a noun), then grammar requires it to be grammatically singular: we must say 'the nearest five is...', not 'the nearest five are...'. In contrast, however, 'five' applied as an adjective to a certain fivesome of cows, say, must be in the plural, qualifying a (perhaps understood) plural noun: we must say 'these five (cows) are...', not 'this five (cows) is...'.[20] And again, when 'five' is used as a tally in counting *F*s, to record that five *F*s have been counted so far, it carries a strong suggestion of multiplicity rather than unity.

The dissonance here, between the firmly plural adjective and implicitly plural tally, on the one hand, and the firmly singular noun, on the other, is enough by itself to generate uncertainty in our minds about using the term as a singular noun to refer some particular fivesome. But in addition, out of these three basic ways of using 'five' in relation to (say) a particular plurality of cows, the use as a count noun to refer to the plurality is clearly derivative from and dependent on the other two uses.[21] And when we are inclined to use it as a singular noun, referring to the cows as 'this five', we shall tend to be swayed, whether consciously or not, by some sort of awareness of the links this use has with the other two, and its dependence on them—an awareness that what we refer to is a fivesome in virtue of being five cows (where we may also have used 'five' as a tally in counting them). As a result, we shall tend to wonder whether 'five' in this use as a noun ought perhaps to be in the plural also.[22] Grammar, however, forbids this; and it is a sense, however dim, of this clash which makes us feel uneasy about using 'five' as a noun to refer to a particular fivesome at all.

---

[19] These possibilities, along with others, are discussed more fully in "Plural Reference," §§ 2 and 7.

[20] A related awkwardness is the grammatical requirement to say, for example, 'A number of men is [rather than 'are'] waiting to see you', despite its strangeness: we clearly think of the expression 'a number of *F*s' as serving to refer, not to some (unspecified) cardinal number, but to a plurality of *F*s (a number-of-*F*s—it is equivalent in function to 'some *F*s'), yet grammar pushes us to couple it with a singular verb. (One case where linguistic practice is currently tending to override the demands of grammar, incidentally, is the noticeable trend in contemporary British usage, perhaps following that in North America, to use a plural verb with words like 'committee' or 'team'.)

The use of 'five' as a singular noun with type reference, in contrast, is not similarly inhibited by these influences, mainly because it occurs within statements that convey universal facts about counting and the results of acts of counting. Here its use as a noun is less obviously dependent on its applicability as a plural adjective and its use in any particular counting.

How, finally, if number adjectives are plural and tallies imply plurality, could a use of number words as singular nouns ever have evolved? Regardless of grammar, we do recognize that a fivesome of *F*s, though it is many relative to one principle of counting, is also one thing relative to another principle; and there is a clear role in language for a singular expression to refer to it explicitly in this latter guise—the role for which we have been using the term 'fivesome', and for which terms like 'quintet' have been coined.[23] But for the reason indicated, when 'five' is used for this purpose, there is a restraint on its use with particular reference, which does not extend to its use as a type term.

---

[21] Hurford (*op. cit.*, §§ 5.1-5.4) argues from a linguistic perspective for the primacy of the adjective over the noun.

[22] Language users unaware of the grammatical descriptions employed in the course of this explanation or unable to apply them clearly may still feel something of the grammatical tension just identified. The uncertainty will be the greater in a language such as English, which uses the same words, or homonyms, as cardinal number adjectives, tallies, and nouns. But there will be a similar if perhaps weaker influence even in a language where this is not the case, if a singular/plural dichotomy is embedded in its grammar—as seems likely to be the case wherever the attribution of numerical size on the basis of principles of individuation and counting is well established. The divergence between grammar and our thinking is recognized, for example, by Simons (*op. cit.*, § 4.3), who distinguishes between syntactic and semantic singularity or plurality.

[23] Simon (*op. cit.*, § 4.4) identifies phrases like 'the class of' or 'the collection of' as another device with the same function. We can conceive of various possible ways

An explanation along these lines, I believe, accounts for our not using number nouns with particular reference as commonly as we might expect of terms usable as type terms, and thereby in turn for our not recognizing number talk as a form of type talk. That said, however, this nonrecognition needs to be properly acknowledged: given that meaning has to be seen as determined by what we understand it to be, we cannot say simply and without qualification that our talk about numbers is a variety of talk in terms of types. The claim will have to be, rather, that this talk, and the corresponding thinking, do in fact function in the way in which talk and thinking framed in terms of category-straddling types generally do, but they are not generally understood as such, and are in fact generally misunderstood as being of some other kind (with considerable confusion and uncertainty about what this kind is—whence the various unsatisfactory attempts to say "what numbers are").

IX. EXTENDING THE ANALYSIS

Can this type analysis of our everyday talk and thinking about numbers be extended convincingly to all talk of numbers within mathematics? There are two issues here. First, is there anything in the formal version of the arithmetic of the integers which cannot be accommodated by the type analysis? Secondly, can the analysis be extended from the integers to other kinds of numbers—rational and real numbers, complex numbers, transfinite numbers? Let me offer some remarks here about the first issue, and then look briefly at the second in section x.

The main question about extending the type analysis to the formal arithmetic of the integers is about universal or existential quantification over numbers, in statements such as:

(F1) For any integers m and n, $m + n = n + m$.

---

in which existing number words like 'five' could have come to be employed for this purpose. One obvious thought is that failure to distinguish between that which 'five' supposedly denotes and the numeral '5' can allow 'five' to be thought of, confusedly, as a (singular) name, with the question whether it denotes the numeral or something else left in obscurity. Another, wholly speculative, suggestion: a folk-etymological interpretation of our use of 'five' as a tally in counting might first take it as a noun (because it appears to stand on its own), and then, more specifically, take it to be a singular noun denoting the type instantiated in the collection of individuals counted up to that stage in the counting procedure. (Either account might explain, incidentally, why we use 'five' on its own as a type term without the prefixed definite article.)

(F2) There is a prime number between 5 and 10.

Universally quantified statements about integers present no difficulty for the analysis of number talk as a kind of type talk. (F1), for example, asserts the commutativity of addition (already mentioned briefly in section VII (7)). It has a content similar in kind to statements such as '5 + 3 = 3 + 5', but with a wider universality: what the latter says about the results of the two different ways of counting the union of any two mutually disjoint pluralities of the specified sizes, it asserts for any two pluralities of any two sizes whatever—much as saying, for example, 'All the big cats are carnivorous' (meaning that each of the species in question is carnivorous) is saying that any (normal) instance of any of these species is carnivorous. Unlike this latter statement, however, (F1), with its truth rooted in facts about the nature of counting, is not contingent but necessary; it holds necessarily for any two plurality types whatever, irrespective of whether they happen to have any instances, and if there are any instances, will apply necessarily to all of them rather than merely happening to be true of all of them. And similarly for any other true statements holding for all integers.

What of existential quantification? Bearing in mind the special noncontingent status of statements about integers, we can appreciate that asserting the existence of a plurality type satisfying a certain condition, as in (F2), is in effect affirming the possibility of there being a particular plurality that would on counting turn out to be an instance of a plurality type of that kind—that is, affirming the possibility that a particular plurality might turn out on being counted to be of such a numerical size that it was both intermediate in size between a fivesome and a tensome, and also prime (not admitting of arrangement without remainder as an m-some of n-somes, unless either the m-some or the n-some was a onesome).

It is clear that in arithmetic we do understand existential quantification over the integers as in effect serving to affirm possibilities regarding particular pluralities of the relevant cardinalities. We do not regard a statement like (F2) as merely contingently true, or accept that arithmetically provable statements about the existence of very large integers might in fact be false for lack of the relevant instances. Reconciling this noncontingent status of existence statements in arithmetic with their having application to the contingent world of our experience is a familiar problem for any nonempiricist account of arithmetic; and the recourse, as here, to a modal analysis of such statements is a common way of dealing with that problem. While a modal analysis fits in easily with our ordinary understanding of both universal and existential statements in arithmetic, its adoption will be defensible only if it makes clear both what kind of possibility is in question, and also how we can know about that kind of possibility. The account just offered does readily meet both of these requirements: the possibility it refers to, possibility relative to the rules of the procedure of counting and the conditions for its applicability, is of a quite ordinary and well-understood kind; and our knowledge of that possibility it presents as being derived from our grasp of these rules and conditions.[24]

Statements like (F2) are in fact rather like assertions about possible moves in chess—for example, an assertion that from a certain possible position, there is a way in which White can win in two moves, no matter how Black responds. Such a way exists, if it does, as a possibility implicit in the rules of chess, irrespective of whether any player has ever actually made the moves in question; and we can deduce its existence from our knowledge of these rules. Rather similarly, (F2) in effect asserts that there is a counting (that is, counting type) such that if it were instantiated in the counting of some particular plurality, then that plurality would (necessarily) satisfy the condition in question.

Statements involving multiple quantification, though their analyses will obviously be more complex, do not present any new problems. Nor, as we have already seen, do functional expressions, such as 'the square of...'. There is every prospect, then, that the analysis of numbers as types can be extended without difficulty to the formal arithmetic of the integers.

X. TALK ABOUT OTHER KINDS OF NUMBERS

Will the analysis extend to talk about other kinds of numbers? The following, necessarily sketchy remarks suggest a positive answer about rational and real numbers; complex and transfinite numbers are rather different cases.

(i) *Rational and real numbers.* The obvious progression from linking integers with counting leads us to think of rational numbers as connected with measuring. Having recognized three as a plurality type having particular threesomes as its instances, we move naturally to seeing a nonintegral number such as 2¼ or 1.7 as a *measured-quantity type*, one whose instances will be particular measured quantities—particular quantities of some kind of measurable "stuff," *taken relative to some unit of measurement*: the 2¼ liters of milk in this bottle, or the 1.7 kilometers which separate my house from the university.[25] Where an instance of three is an aggregate taken relative to a principle of counting, on this analysis an instance of 2¼ is a concrete quantity of a kind that admits of measurement, taken relative to some measurement unit. Any such instance is, on the one hand, an object of experience, but, on the other, it is that object taken in a certain way, relating it to a measurement unit; and this means that 2¼, the rational number qua *type*, corresponds to a ratio or proportion, the ratio of comparative magnitude which holds between any of its instances and the unit quantity to which we are relating that quantity in thinking about its magnitude. (Recall here the ancient Greeks' treatment of rational numbers as ratios.)

---

[24] On the view that numbers are design types, the problem of contingency just discussed would of course not arise, since a design type may exist, as a cultural object or object of thought, without ever actually being realized in a particular token (recall again the never-performed opera). The account of number talk offered here does not adopt this view; but it does present a statement like (F2) as closely linked with an assertion about a counting type, which is something akin to a design type in that its existence as something instantiable is determined solely by what the rules of counting and conditions of countability require, irrespective of whether it is ever actually instantiated.

This account explains very easily how statements about rational numbers apply to our experience, without any need to reify ratios (it requires us to recognize only procedures of measuring relative to a unit, particular quantities that admit of being measured, and the corresponding measured-quantity types). And while standard rational-number theory, framed in terms of ordered pairs of integers that satisfy certain axioms, may appear very remote from statements about measured-quantity types, it can readily (as one possible interpretation, another being as a theory of ratios) be linked with such statements once we bring into play three things: the analysis of integers as plurality types; the recognition that measuring relative to a unit of measurement involves in one way or another counting complete units and complete subunits (halves, quarters, tenths,...) of the magnitude in question; and the necessary general conditions for any kind of quantity to admit of being measured.[26]

---

[25] Strictly, being an instance does not involve actually being measured, just as being an instance of seven or the sevensome does not involve actually being counted. What is essential is that the quantity is taken relative to some unit of measurement.

[26] Of these conditions, the main ones relate to certain sorts of stability—invariance, or fluctuation within manageable limits—which the particular instances of the kind of variable quantity in question must exhibit; to the existence of means by which we can satisfy ourselves that two particular quantities of that kind are equal

What about irrational numbers? A certain measurable quantity (the length of the diagonal of a square, say) may fail to stand in any exact ratio (the ratio between two integers), to a given unit of measurement (for example, the length of a side of that square): it and that unit are incommensurable. Accepting such facts—which are universal negative facts, and not actually discoverable from experience of instances but only from theory—forces us to recognize irrational numbers alongside rational ones. An irrational number is a type any instance of which is, like an instance of a rational number, a particular measurable quantity taken relative to some unit of measurement, but one where the quantity in question and the particular unit happen to be incommensurable. Irrationality is always relative to the unit of measurement employed, and may be eliminated by a change of unit. Thus there is nothing mysterious about an instance of an irrational number. And when we talk about a real number, which may be either rational or irrational, we are talking about a *measurable-quantity type*, each of whose instances is a particular measurable quantity taken relative to some unit of measurement. Where measurable quantity and unit are commensurable, the real number is rational, a measured-quantity type; where they are not, it is irrational.[27]

Real numbers, so understood, may seem very unlike the very abstract entities (infinite sets, pairs of infinite sets, or convergent infinite sequences of rational numbers) with which real-number theory standardly identifies them. But that theory should admit of interpretation as a theory of measurable-quantity types, by exploiting the conceptual link between a convergent infinite sequence of rationals and a idealized process of approximating ever more closely, given ever more accurate means of measurement, to the precise measurement of some particular quantity, taken relative to a chosen unit. (This interpretation will not, of course, exclude other possible interpretations of the theory, taken as an abstract axiomatic theory; but it meets the pressing challenge facing real-number theory of explaining how that theory can have application to measurable magnitudes in our experience.)

---

in magnitude; and to constancy in results (the measurements arrived at being invariant under different ways of applying the measuring procedure).

[27] This account has the major virtue of making the rationals a subset of the reals, whereas any of the alternative standard definitions of a real number in real-number theory defines it in terms of an infinite set or sequence of rational numbers, thereby treating it, counterintuitively, as something quite different in kind from a rational number. It also avoids the charge of arbitrariness which can be leveled against choosing any one of the different ways on offer for defining the reals in terms of the rationals.

(iii) *Transfinite numbers; complex numbers.* The instantiations, if any, of a transfinite cardinal will by definition not be identifiable as such by counting. But given two uncountable pluralities (whether real existents or only mental constructs), we may sometimes by some other means satisfy ourselves that they are equinumerous (usually by establishing a one-one correspondence), or that one is larger than the other (usually by reductio arguments, including diagonal arguments). On such reasonings transfinite arithmetic has been built up. And a transfinite cardinal may be taken as a type whose instances, if it does have any, will be uncountable (perhaps nondenumerable) but equinumerous pluralities. (If it has no instances, whether as a matter of fact or of necessity, then theorems about it will, in effect, be necessarily true universally quantified conditional statements, the antecedents in which are—in fact or necessarily—never satisfied.)

In contrast, a complex number is such that no single object of our experience could possibly be seen as an instance to which it stands as type. One response to this fact, I suggest, is to conclude that complex-number theory cannot be seen as dealing with another special kind of numbers alongside but distinct from the integers, the rationals and the reals; and that a complex number needs rather to be taken as a matched-pair type whose instances are *matched pairs of measurable quantities*: quantities that exhibit distinctive relationships, where it turns out that these relationships can be captured using the concepts of the special kinds of addition and multiplication characteristic of complex numbers (examples being certain quantities studied in electromagnetic theory and in quantum theory). But that suggestion cannot be pursued here.

XI. SOME IMPLICATIONS: SUMMING UP

If the type analysis of the form our thoughts and statements about numbers take is correct, what, if anything, does it imply about the content of these thoughts and assertions? And how does it relate to other accounts of numerical mathematics? Its concern, being only with the vehicle of our thought, might seem rather superficial, but it can shed some helpful light and dispel some misunderstandings. For simplicity, let us confine our attention in this brief discussion to the case of talk about the integers.

(i) First, does the analysis support a realist or a reductivist view of our apparent references to numbers? It could be interpreted in either way, according as talk about the *n*-some is taken as referring to something real (presumably a universal of a distinctive kind), or as merely a convenient shorthand way of saying things about all particular *n*-somes of every kind. In fact, however, on the type analysis the question about the reality of numbers becomes just one special case of a wider question about the status of types generally; and this question does not fall to be answered specifically within the philosophy of mathematics. Indeed, it does not seem necessary to resolve either this wider question or the narrower one about numbers as types in order to answer any other major philosophical questions about numerical mathematics. If the account offered above of how the truths of arithmetic acquire their necessity, epistemic certainty, and applicability to experience is accepted, we can understand the nature of numerical mathematics and how we apply it to the world while suspending judgment on the status of types in general, and numbers as types in particular.[28]

(ii) *Structuralism.* In fact, the existence of multiple interpretations for a formalized version of arithmetic, including some which are not obviously numerical,[29] makes it hard to maintain a realist view about numbers, and strongly supports some sort of structuralist account of numerical mathematics—with which, as remarked already, the type analysis meshes very naturally. Any such account faces two questions: (1) How is it possible for a theory apparently framed in terms of things of one particular kind—in this case, numbers—and their particular properties and relations to be really or in effect about a general pattern of linked properties and relations, a pattern which may be exemplified in the particular properties and relations of things of diverse kinds and even different categories? And (2) how can structuralism present the statements of pure-number theory as being "in effect" about such a general pattern or structure, without then having to reify that abstract structure as an entity in its own right?[30]

The type analysis provides neat and simple answers to both questions. The category-straddling nature of numbers as types explains, in answer to (1), how we are able to speak simultaneously about things of very different kinds in the convenient shorthand of statements about numbers, as types that have things of each of these kinds as instances. This is a simpler and more natural explanation than one framed in terms of the possibility of providing different interpretations for the formulas of a theory that is constructed initially as a purely formal and uninterpreted system or calculus. And in considering (2), once we recognize the direct link between numbers as types (and the properties and relations we attribute to them) and the particular pluralities of different kinds that are their instances (with their corresponding properties and relations), we can see that there is no need to reify an additional "third thing," in the form of an abstract pattern of interconnected properties and relations which would constitute the true subject matter of arithmetic: the things we say about numbers apply to each kind of instance immediately, and not through any abstract intermediary.

---

[28] Any realist view of types which treats them as genuine universals has in fact to handle two difficulties. The first is the fact that we regard types, unlike other universals, as capable of having some of the properties their instances can have. The second is that of explaining why we should recognize the existence of two kinds of universal, the type (the F) and the corresponding sortal property (F-ness), when the latter appears to be more basic, and anything expressible in terms of the type can apparently be reformulated in terms of it.

[29] See, for example, Bigelow's (*op. cit.*, chapters 11-12) ingenious interpretations in terms of relations holding between types of genealogical relationships, in the course of developing the Whitehead-Wiener-Quine conception of mathematics.

[30] Parsons, in "The Structuralist View of Mathematical Objects," p. 311, identifies this as a central challenge for structuralism.

A further point. Structuralism developed initially in part from awareness of the arbitrariness of identifying the natural-number sequence with any one of various theoretically defined sequences (for example, with the Zermelo sequence Λ, {Λ}, {{Λ}}, ... , rather than the Von Neumann sequence Λ, {Λ}, {Λ, {Λ}}, {Λ, {Λ}, {Λ,{Λ}}}, ...), even though any one of these identifications would yield the right truths about the natural numbers.[31] The arbitrariness here stems from the incompletely determined nature of numbers (and hence of the number sequence); and that, in turn, is readily traceable to a number's being a category-straddling type, and having no characteristics except those which every one of its categorially diverse instances has. (Similarly, a word type has an incompletely determined nature.) Any instance of the sequence type (and this includes the Zermelo and Von Neumann sequences themselves, despite their also being types) will have just the right relations holding among its members; but it will also have some other properties and relations peculiar to itself. The only nonarbitrary identification, and the correct one, is with *the sequence type of which they all are instances*: this alone has just those features we attribute to the number sequence—precisely those features and only those features which every one of the instances has. This conclusion provides a defensible and nonarbitrary answer to the question what the number sequence is, while also incorporating the structuralist insight.

Finally, notice that what has been said about structuralism and the type analysis extends straightforwardly also to the nonnumerical parts of mathematics (where, once again, the existence of multiple interpretations of theories supports a structuralist account). The type analysis will apply just as naturally to talk about sets, groups, lattices, matrices, and so on as it does to talk about numbers; and in each case it answers the two important questions just identified, thereby providing important collateral support for a structuralist analysis.

---

[31] See Paul Benacerraf, "What Numbers Could Not Be," *Philosophical Review*, LXXIV, 1 (January 1965): 47-73, especially §§ I-II.

(iii) *Field's instrumentalism.* The central thesis of Hartry Field's *Science without Numbers*[32] is that we can understand and justify the application of mathematics in a given science without having to regard the statements of numerical mathematics as true or false—and so without having to take the references to numbers in them as genuinely referring to real entities—provided that we can show that adding the statements in question to the body of statements accepted in that science yields a *conservative extension* of that body, that is, that adding these resources does not enable us to deduce anything that could not be deduced without that addition. We may properly use arithmetical statements, without regarding them as anything more than reliable fictions, so long as we know that we shall not thereby be led astray, in that using them will yield only conclusions that we could establish independently (and hence reliably) without them. The use of arithmetic is justified purely by the fact that it is a reliable (and, in principle, dispensible) instrument, not by the truth of its statements.

The type analysis, in contrast, offers a simpler and more direct justification of the use of arithmetical statements in science, by presenting these statements as the straightforwardly meaningful and true statements we naturally take them to be, while at the same time satisfying Field's wish to avoid taking them as referring to any abstract entities of an obscure kind. By analyzing them as statements effectively about plurality types, it warrants our use of them in passing from one set of factual statements about particular pluralities of a given sort or sorts to further factual statements of the same kind. Once this simpler analysis and justification are on offer, Field's revisionary account of our understanding of the status of arithmetical statements appears not to be needed in order to avoid accepting the existence of numbers as real abstract entities of a peculiar kind.

---

[32] Princeton: University Press, 1980, especially pp. 1-23. While suggesting below that there is a simpler way of satisfying Field's desire to justify the use of numerical mathematics in science without reifying numbers, I regard his detailed and sophisticated analysis of concepts central to physics as very valuable in other ways. Further, there is a role for some kind of instrumentalist approach in analyzing certain techniques in various areas of mathematics and formal logic, and more generally in understanding the nature and role of calculation in mathematics. Also, Field's approach might provide another possible way of making it intelligible how complex-number theory can have application to reality. (Notice, incidentally, that it is true on the type analysis, as it is on Field's, that in our dealings with the world we could dispense with arithmetical truths: we could replace them with universal statements about pluralities.)

(iv) *Accounts of numbers as properties, relations, or operations.* Various other recent writers have felt the pressure to say "what a number is" or "what arithmetical statements are about" (even though most of them do not favor a view of numbers as real objects, many of them are following in the empiricist tradition of J.S. Mill, amended to accommodate sortal relativity, and some have structuralist sympathies). David Armstrong identifies the number n with the property of being n-parted. John Bigelow identifies the number five, for example, with the relation that holds among any five distinct things, the relation of fivefold mutual distinctness. Glenn Kessler makes allowance for sortal relativity by identifying a number with a relation that may hold between an aggregate and an "individuating property." Philip Kitcher (influenced by Mill's empiricist account of our arithmetical knowledge as acquired through our actions) offers a reformulation of arithmetical statements as statements about relations holding between arithmetical operations: the operations of singling out one thing, singling out two distinct things, and so on.[33]

Making no more attempt than in the remarks about Field to do justice to the rich and often illuminating subtleties of the overall accounts within which these identifications appear, let me merely suggest that they all diverge from our ordinary understanding of number words, which takes these words as apparently referring to objects of some kind, rather than to properties of objects, or to relations holding between or among objects, or to types of action; and that the divergences are unnecessary.

The divergences come out in the complex interpretations the accounts will have to give of a statement such as 'Five is greater than three'. If, for example, 'five' and 'three' refer to what we may call

---

[33] See, respectively, Armstrong, *Universals and Scientific Realism* (New York: Cambridge, 1978), Volume II, chapter 18, § 5; Bigelow, chapter 6; Kessler, "Frege, Mill and the Foundations of Arithmetic," this JOURNAL, LXXVII, 2 (February 1982): 65-79; Kitcher, *The Nature of Mathematical Knowledge* (New York: Oxford, 1984), chapter 6, § II; I am grateful to an editor of this Journal for the reference to Kessler. (Leaving out the identifications referred to in the text, much of the rest of what these accounts say about numerical mathematics could readily, with appropriate modification, be reconciled with the type analysis. In regard to Kessler's account specifically, the point to be made is that a reference to a plurality of particulars already associates an "individuating property" with that plurality's constituents, and hence with the plurality itself: the relativizing is already built in to the thing to which we attribute cardinality, so we do not need to treat what is attributed as being itself a relation.)

"five-partedness" and "three-partedness," then the statement is naturally read as asserting that five-partedness is a greater property than three-partedness. But what (if anything) does 'is greater than' mean in relation to properties? Identifying numbers with relations similarly makes it difficult to interpret such statements in any simple way. Kitcher's interestingly different account, which perhaps avoids this difficulty, would serve exactly as an account of the counting types (act types of counting) corresponding to the natural numbers (though he nowhere actually mentions counting). His reformulation of arithmetic, however, seems to substitute the process (counting up to *n F*s) for what is identified through that process (the plurality discovered to be an *n*-some).

Each of these accounts may well deliver the goods formally, yielding as consequences the standard truths of arithmetic. But each involves a revisionary understanding of the kind of meaning we actually conceive arithmetical statements to have, and a more complex account of how they apply to the world, where the revision and complexity are not needed, given that the type analysis yields an account adhering very closely to our ordinary understanding of their meaning and providing a very simple account of how they apply to reality, without reifying numbers as some quite special kind of entity.

(v) *Summing up.* If the challenge discussed in section VIII has been deflected, then the thesis that our everyday talk and thinking about numbers—and talk and thinking in arithmetic about the integers, the rationals, and the reals—are framed in terms of types (although not so understood by us) must surely commend itself strongly for acceptance.

Indeed, while the type analysis has been offered as an account of how our ordinary talk and thinking about the integers does actually function (unrecognized by us), even someone who is not convinced by it as such a descriptive analysis would still, I suggest, have good reason to accept the account, taking it instead as (from their point of view) a mildly revisionary analysis, suggesting how our talk and thinking should be re-interpreted. For the analysis does fit and account for the facts about number talk, in a simpler and more economical way than any alternative account. In particular:

(1) It gives a simple answer to the very basic question about the applicability of numerical mathematics. And by bringing out in the course of this answer the links between our dealings with such collections and the procedures of counting and measuring, it points the way to an explanation of how the truths of arithmetic can have a special kind of necessity, and how we can know these truths with certainty.

(2) As indicated in (i) above, it provides a simple answer to the question 'What are numbers?', which frees the analysis of mathematics and how it functions from any need to wrestle with ontological questions that may be beyond resolution.

Finally, (3) the analysis and these answers do not invoke anything other than three wholly familiar components of our everyday speech and thought, namely: the concept of a plurality, the kind of thing we refer to whenever we use a plural-count noun phrase referringly (and, for the analysis of talk about rational and real numbers, the corresponding concepts of a measured and a measurable quantity); the pattern of making general (in this case, straightforwardly universal) statements using type terms with distributive reference; and the procedures of counting by reference to a principle of individuation and counting, and of measuring quantities relative to some standard unit quantity.[34]

J. R. CAMERON

University of Aberdeen

---

[34] What has not been provided here, for reasons of space, is any formalization of the type analysis to compare with the formal treatments in the accounts discussed in (iii)-(iv) above. That lack is not difficult to make good, and I hope to remedy it elsewhere.

TWO STANDPOINTS AND THE BELIEF IN FREEDOM*

Immanuel Kant[1] poses a problem. We must think of ourselves as free in order to "make use of reason in our conduct" (*ibid.*, p. 56 [GW455-56]). Yet by reasoning about the very possibility of experience, we find that everything that happens is caused to happen, of necessity, in accordance with the laws of nature. Since freedom is a property of the will of a rational being "that makes it effective independent of any determination by alien causes," reason appears to be the source of recognizably contradictory beliefs. As Kant writes, "there arises [...] a dialectic of reason, since the freedom attributed to the will seems to contradict the necessity of nature" (*ibid.*, p. 56 [GW455]). Reason seems to yield the belief that one is free and that one's actions are undetermined, on the one hand, and the belief that one's actions are determined and that one is therefore not free, on the other.[2] How can this be explained?

Kant's own responses to this question are both tantalizing and notoriously difficult to understand. As a result, a great variety of interpretations abound. Until recently, different versions of the *two-worlds account* and the *two-aspects account* dominated the interpretive debate. According to these accounts, contrary to initial appearances, reason does not in fact commit us to contradictory beliefs. For this reason, both accounts fall into what I call the *noncontradictory contents* category of solutions to Kant's problem.

Here their similarity ends. On the two-worlds account, reason leads us to two beliefs about two different subjects or worlds. For example, our belief that we are free and our actions undetermined is a belief about our noumenal selves, and our belief that our actions are determined is a belief about our phenomenal selves. Thus, we can remain perfectly consistent and rational in believing the deliverances of reason.

In contrast, on the two-aspects account, our beliefs are about the same world and about the same objects; but one of our beliefs is

---

* I am very grateful to Tyler Burge, Barbara Herman, Pat Matthews, Sam Rickless, and Carol Voeller, all of whom read previous drafts of this paper and provided especially helpful comments.

[1] *Grounding for the Metaphysics of Morals*, James W. Ellington, trans. (Indianapolis: Hackett, 1981, second edition). Also see p. 50 [*Gesammelte Werke* 448]: "Now I claim that we must necessarily attribute to every rational being who has a will also the idea of freedom, under which only can such a being act."

[2] I shall use 'determined' throughout as shorthand for "determined by natural causes in accordance with natural laws."

0022-362X/00/9710/564–76

about ourselves and our actions considered in one aspect or under one description, while another is about ourselves and our actions considered in a different aspect or under a different description. Consider, for example, a version of the two-aspects account based on Donald Davidson's[3] anomalous monism. When an event is described as a mental event or an action, it cannot be predicted or explained by physical laws and prior events. Yet that same event, when described in physical terms, can be predicted and explained in terms of physical laws and prior events. Thus, it is perfectly consistent that an event, under one description, falls under strict deterministic laws, and that the same event, under another description, is free of the "nomological net." Further, it would seem that one can be perfectly rational in believing that one's actions are free (when they are described in one way) and that they are not free (when they are described in another way).

Yet both accounts face serious difficulties. For example, the two-worlds account faces the pressing problem of how to explain the relationship between the two worlds to which it appeals. It seems that the "noumenal I" must bear some relation to some of the actions that take place in the phenomenal world of time and space. Otherwise, the fact that the "noumenal I" is free and its actions undetermined would have no relevance to the actions for which we normally take people to be responsible. Yet it is unclear how anything in the noumenal world, distinct from the phenomenal world of space and time, could "affect" that world. Thus, the two-worlds view saddles us with a metaphysical apparatus that seems to raise as many problems as it solves.

The two-aspects account avoids these problems, but has problems of its own. For example, it has often been noted that, if an event is determined, then it is determined no matter how it is described. For this reason, Davidson's version fails to provide a reconciliation of free action (understood as undetermined action) with determinism. Nor can anomalous monism be used to exonerate reason for producing apparently contradictory beliefs. For as long as one can become aware by reasoning that one's actions, considered as mental events, are the same things as one's actions, considered as physical behavior, then one is irrational in believing that one's actions are undetermined and one's physical behavior determined. Presumably, we can become aware of the identity of our actions and our "determined" physical behavior; otherwise, we could not understand the two-aspects view as offering a solution to the problem posed by Kant.

---

[3] "Mental Events," in *Essays on Actions and Events* (New York: Oxford, 1980), pp. 207-27. Davidson suggests that his view provides a solution to the problem posed by Kant, but does not claim to provide a rigorous interpretation of Kant's own solution to it. But for an attempt to defend anomalous monism as a faithful interpretation of Kant, see Ralf Meerbote, "Kant on the Non-Determinate Character of Human Actions," in W. Harper and R. Meerbote, eds., *Kant on Causality, Freedom and Objectivity* (Minneapolis: Minnesota UP, 1984), pp. 138-63.

Replies to these objections have been made on behalf of both the two-worlds and the two-aspects accounts, and subtle versions of each have been developed.[4] At the same time, finding a plausible alternative is an attractive prospect.

In fact, a distinct kind of account has begun to emerge, an account according to which there is nothing problematic in believing both that our actions are free and so undetermined, and that our actions are determined and so not free, because each of our beliefs is held from different standpoints. In other words, although reason commits us to contradictory beliefs, there need not be any irrationality as a result, as long as each belief is held from a distinct standpoint. I call this the *two-standpoints account.*[5] I explore here some of the ways in which the two-standpoints account has been and can be developed, and I assess the results. I do not (directly) evaluate the account for its faithfulness to Kant's text, but instead assess it for its internal coherence and for its ability to solve the problems it is intended to solve.[6] I argue that while it has certain advantages over its rivals, it does not ultimately succeed.

---

[4] For example, Henry Allison claims that his version of the two-aspects account avoids the problems described in the text which afflict anomalous monism—*Kant's Theory of Freedom* (New York: Cambridge, 1990), especially pp. 43-46. Others have also offered another sort of noncontradictory account, according to which practical reason commits us to the belief that we are free in one sense while theoretical reason commits us (or might commit us) to the belief that we are unfree in a different sense. For example, Hilary Bok, who does not claim to be providing an interpretation of Kant's own solution to his problem, offers a version of this account in *Freedom and Responsibility* (Princeton: University Press, 1998). While this kind of account is a promising one, it faces the challenge of explaining why skeptics about freedom themselves assume that their skeptical arguments cast doubt on the very same notion of freedom that they presuppose when faced with decisions about how to act. See, for example, Galen Strawson, who argues in *Freedom and Belief* (New York: Oxford, 1986, pp. 25-31 and pp. 69-70) that human beings who deliberate about how to act cannot help believing themselves to be free in the same sense of 'free' that he has shown by philosophical argument to be impossible to instantiate.

[5] This terminology brings with it a source of potential confusion: sometimes noncontradictory contents accounts are described as identifying "two points of view" from which the same object can be considered. For example, in *Kant's Theory of Freedom,* Allison explicitly characterizes his view as a version of the "two-aspects" interpretation of Kant's transcendental idealism (p. 4). He then goes on to say that his account "offers instead the [...] contrast between two 'points of view' *or* descriptions under which a single occurrence (a human action) can be considered" (p. 4, emphasis mine). Similarly, in *Freedom and Responsibility,* Bok offers a noncontradictory contents account, arguing that practical reason gives us reason to believe ourselves free in a compatibilist sense of 'free', while theoretical reason (can) give us reason to believe that we are not free in an incompatibilist sense of 'free' (see pp. 120-22). At the same time, she refers to her argument as a "standpoint argument" (see p. 52), and certain passages in which she summarizes her view can be read in a way such that the beliefs in question appear to be contradictory after all (see pp. 160-62). To avoid confusion, I shall reserve "two standpoints" for views which focus *exclusively* on the role of distinct standpoints to address the problems set by Kant,

I. THE TWO-STANDPOINTS ACCOUNT

According to the two-standpoints account, the propositions to which reason commits us are indeed contradictory. But we are not irrational in believing that we are free and undetermined, on the one hand, and believing that we are determined and so unfree, on the other, because we hold the apparently contradictory beliefs from different standpoints. For example, it is from the deliberative standpoint that we believe that we are free and our actions undetermined, while it is from the theoretical or scientific standpoint that we believe that every event (including our actions) is caused by previous natural events in accordance with the laws of nature, and so determined. This account differs from the two-worlds account in that the beliefs in question are about the same world and the same objects. It also differs from the two-aspects account in that the subjects of the relevant beliefs are considered under the same aspect or description.

Christine Korsgaard,[7] for one, suggests something like this view when she interprets Kant's claim that practical reasoners, or rational beings who have a will, must act under the idea of freedom. She writes

---

and "noncontradictory contents accounts" for views that focus primarily on the role of distinct contents in the beliefs yielded by reason.

[6] One reason for considering two-standpoints accounts on their own merits is that they are of interest not only to interpreters and defenders of Kant, but also to those who accept only some of the premises of the problem posed by Kant. For example, there are those who accept that we must believe ourselves free in order to reason practically, and that—although reason itself does not demand it—we might come to be justified in believing that we are not free on the basis of evidence—see, for example, Peter Van Inwagen, *An Essay on Free Will* (New York: Oxford, 1983), pp. 153-61. Since both of the apparently contradictory beliefs are not the products of reason alone, this is not exactly Kant's problem. Nevertheless, a similar problem arises. Should we say that a practical reasoner who came to believe on the basis of evidence that she is not free is irrational? Two-standpoints accounts provide a way of answering this question in the negative.

[7] "Morality as Freedom," in Yirmiyahu Yovel, ed., *Kant's Practical Philosophy Reconsidered* (Boston: Kluwer, 1989), pp. 23-48. Elsewhere in the same article, Korsgaard appears to endorse a very different reading of Kant, according to which acting under the idea of freedom does not entail that one has a belief. For example, in explaining Kant's claim that in making a choice, one must act under the idea of freedom, Korsgaard writes, "[t]he point is not that you must *believe* that you are free, but that you must choose *as if* you were free" (p. 26). I am not sure how to reconcile the passage quoted in the text with this one.

that "[t]he standpoint from which you adopt the belief in freedom is that of the deliberating agent" (*ibid.*, p. 38). There need not be any irrationality in simultaneously possessing the belief that one is determined and the belief that one is free and so not determined, however, because the beliefs are held from different standpoints. They are about the same world, but they view it differently, and in such a way that they do not conflict.[8]

Thomas Hill[9] appears to offer a similar account when he writes that from the "*point of view of deliberation*" one sees oneself as having options, and that "whether one's reflection and choice are themselves causally determined, even whether the agent believes this *as a metaphysical thesis*, is irrelevant to his deliberative task" (*ibid.*, p. 371). Although Hill does not here put the point in terms of "freedom," he offers an account that is structurally similar to Korsgaard's: one can believe *p* (in this case, that one has options) from the deliberative point of view, and one's metaphysical belief to the contrary is "irrelevant."

Now we can ask how it is that this account explains the fact that one can have beliefs toward contrary propositions and yet not be in an irrational state. Normally, to have the belief that *p* and the belief that *not-p* is to be inconsistent. But to have a belief that *p* and, say, the desire that *not-p* is not inconsistent. In the latter case, the attitudes are different. It might be argued that if, in virtue of their being held from different standpoints, the beliefs about freedom are instances of different types of attitudes, then perhaps there is no inconsistency involved in this case either. Advocates of this account do not suggest that having either sort of belief is anything like having a desire. I offer the analogy only in order to show that the account occupies a logical space in which it is not appropriate to assess two beliefs for consistency.

---

[8] See, for example, Korsgaard, "Morality as Freedom": "These two standpoints give us two very different views of the world" (pp. 37-38). "The deliberating agent, employing reason practically, views the world as it were from a noumenal standpoint, as an expression of the wills of God and other rational agents. [...] The theorizing spectator, on the other hand, views the world as phenomena, mechanistic, and fully determined. The interests of morality demand a different conceptual organization of the world than those of theoretical explanation" (p. 37). The reference to conceptual organization raises the possibility that Korsgaard holds a version of the two-aspects view, rather than a two-standpoints view as I have defined it. Although the two-standpoints view is clearly suggested by passages like the one quoted in the text, I am not certain that Korsgaard's view in this article ultimately fits neatly into either category. In *The Sources of Normativity* (New York: Cambridge, 1996), Korsgaard again takes up what seems to be a distinction between two standpoints: "[i]t is from within the deliberative perspective that we see our desires as providing suggestions which we may take or leave" (p. 96). But the "Scientific World View," which "serves the purposes of explanation and prediction," may tell us that certain of our desires are determined to cause particular actions (p. 97). Korsgaard is not here claiming to provide an interpretation of Kant, but is instead endorsing the use of two standpoints to address something like Kant's problem.

[9] "Kant's Theory of Practical Reason," *Monist*, LXX (1989): 363-83; emphasis mine. It should be noted that elsewhere Hill seems to suggest a different reading of Kant, namely, an account according to which one must only assume that one is free for certain purposes—"Kant's Argument for the Rationality of Moral Conduct," *Pacific Philosophical Quarterly*, LXVI (1985): 3-23. It is similar to the account suggested by the passage quoted from Korsgaard's "Morality as Freedom" in footnote 7.

The account explains how we might be epistemically committed to two contrary propositions without being engaged in any sort of irrationality. It also captures the intuitive appeal of the Kantian idea of "two standpoints" without the controversial metaphysics of the two-worlds interpretation of transcendental idealism. Finally, on the two-standpoints account, unlike the two-aspects view described earlier, it is possible both to accept the premises of Kant's problem and to understand the account without being forced to admit to a certain kind of irrationality.

Yet, as I argue, there are several reasons why the account should be rejected. The first is that, despite the appearance of coherence, we might wonder whether "believing from a standpoint" is really an intelligible idea. It is natural to think that belief is an attitude that represents the world as being a certain way—simpliciter. One does not possess such an attitude "from a standpoint."

It is true that we are familiar with the locutions 'standpoint' and 'point of view'. For example, we often say things like, "From one standpoint, it would be best to do *X*, and from another standpoint, it would be best to do *Y*." In this case, what we mean is simply that given certain considerations, it would be best to do *X*, and given other considerations, it would be best to do *Y*. This use of 'standpoint' does not give us reason to recognize the possibility of "believing from a standpoint." Yet there are other examples of familiar "standpoint" talk which at least superficially seem to countenance the idea of believing from a point of view. For example, someone might say: "From the point of view of an American, I believe *p*, and from the point of view of a religious person, I believe *not-p*." What is being said here? I see three alternatives. First, the person might be saying that when she limits her considerations to those concerning the principles underlying the United States Constitution, she concludes *p*, and when she limits her considerations to those concerning the tenets of her religion, she concludes *not-p*. But if this is the correct way of cashing out such talk, then neither her actually believing *p* nor her actually believing *not-p* can be inferred from her statement. Second, she might be undecided about which set of considerations takes precedence, and so is uncertain about which conclusion to draw. Third, she might simply hold two contradictory beliefs, and this fact can be explained without appealing to the notion of "believing from a standpoint." (Admittedly, the plausibility of this last suggestion depends entirely on whether an alternative explanation of the admittedly puzzling situation can be found.) If any of these suggestions captures the correct way of cashing out the "standpoints" talk in this example, then this kind of example should not force us to accept the intelligibility of believing something from a standpoint. And I believe that most of our ordinary "standpoint" talk is best interpreted in ways that do not involve the attribution to people of beliefs from different standpoints.

Let us assume for the sake of argument, however, that we do believe things from different standpoints. The account still faces serious challenges. I begin by considering two together. Both target the ability of the account to explain how one could hold two beliefs toward contrary propositions without conflict. The first concerns the difficulty of producing a criterion that sorts beliefs appropriately into those held from the deliberative standpoint and those held from the theoretical. The second is that, even if such a criterion can be produced, it does not follow that the account will succeed in showing that the existence of the two standpoints is sufficient to prevent irrationality. In what follows, I examine three candidate criteria, each of which purports to distinguish between standpoints.

(A) *The interests criterion.* According to the first suggestion, we believe some things from the point of view of the deliberator and others from the point of view of the scientist or the metaphysician. When we believe things from the point of view of the deliberator, we are engaged in deciding how to act, and when we believe things from the point of view of the scientist, we are engaged in trying to describe and explain the world. Our different sets of interests determine different ways in which we view the world.

In assessing this criterion, we must ask whether the criterion of "interests" succeeds in sorting beliefs appropriately, and, in particular, in such a way that the Kantian contradictory beliefs are held from distinct standpoints. Initially, it appears that the criterion succeeds on both counts. A closer look suggests otherwise, however.

When we are engaged in deliberation, we often rely on theoretical or scientific beliefs. For example, if I am deliberating about whether to sound a fire alarm, one of the things I rely on is my belief about what effects that action is likely to have. Does this mean, then, that my belief about the causal role of alarm sounding is a belief from the standpoint of the deliberator? It certainly seems so, for the belief seems quite "relevant" to my deliberative task. And if so, then it would appear that either I have two beliefs with similar contents that are distinguished by the points of view from which they are held, or I have a single belief that floats freely back and forth between standpoints. In either case, there seems to be nothing in principle that prevents a belief that is held from one standpoint to be held from another.

Recall, though, that the two-standpoints view is supposed to explain the lack of irrationality in holding the beliefs that one is free and that one is not free by relegating the beliefs to different standpoints. If it is to provide such an explanation, then the theoretical belief that one is not free must not be held from the deliberative point of view. But what is to prevent my taking that theoretical belief into account in my deliberation just as I take into account other of my theoretical beliefs? The advocates of this two-standpoints account have the burden of explaining why it is that the theoretical belief that one is determined and not free is "irrelevant to one's deliberative task." They must say why this belief about freedom is not taken up by the deliberative point of view when other paradigmatically theoretical beliefs regularly appear to be. For one's theoretical belief that one is determined and not free would appear to be at least as relevant to one's deliberative tasks as beliefs about the effects of one's actions.

In response, a two-standpoints advocate might argue that the "fire-alarm" example is problematic because it already mixes two sorts of reasoning, each with its own interests, namely, the practical and the theoretical. The part of one's deliberation that is practical in this case is, say, deciding whether or not to save people from a painful death, and this is done simply by consulting the categorical imperative. Practical reason is concerned with decisions about what one ought to do (for example, whether to save people). The question whether one can achieve one's end by some particular means (for example, sounding a fire alarm) is a matter for theoretical reason to decide. Thus, one's belief about the consequences of sounding an alarm is not in fact relevant to one's interests in practical deliberation, and the attempt to show that there is no clear distinction between the relevant beliefs in terms of interests fails.

This is a challenging point. Let me reply by emphasizing an assumption accepted by the two-standpoints advocate: the belief that one is free underlies one's practical reasoning. For this very reason, it would seem that nothing could be more relevant to one's reasoning than one's belief that one is determined and so *not* free. Thus, even if beliefs about the consequences of the soundings of fire alarms or of promise breaking are not employed in one's practical reasoning, it would seem that the belief that one is determined and not free is not only relevant, but of central importance, to that task.

In the absence of further elaboration, then, the attempt to use a difference in interests in order to account for the fact that the belief that one is not free is only held from the scientific standpoint does not succeed. Yet, even if it could be shown that the "interests" criterion appropriately sorts beliefs, an obstacle would still remain for this two-standpoints account.

Earlier, I described the two-standpoints account as making room for the possibility that beliefs from different standpoints, being different kinds of attitudes, might fail to be candidates for conflict. As a familiar example in which this might occur, I mentioned the pair of a belief and a desire. But from this starting point, it is a large leap to the claim that the two kinds of beliefs (distinguished by the two standpoints from which they are held) actually fail to be candidates for conflict. The mere fact of their being attitudes of different types does not guarantee that there could be no inconsistency in holding them.

For example, consider again the religious American. We might say that her beliefs are in conflict even if she holds them from different standpoints, and that rationality dictates that she resolve the conflict. Or consider a case even more closely analogous to the two standpoints at issue: Thomas Nagel's[10] two-standpoints view which recognizes both the "objective" and the "subjective" point of view. Nagel explicitly rejects the claim that having two beliefs from two different standpoints prevents any possible conflict. He claims that from one point of view we believe that we are free, while from another point of view we (can) believe that we are not. But he suggests that, in this case, we have a case of inevitable conflict.

Now, it is true that Nagel's "standpoints" are significantly different from the "standpoints" distinguished on the basis of interests. But the structure of Nagel's view illustrates the fact that merely distinguishing between standpoints does not resolve all concerns about conflict. To conclude that there is no possible conflict between beliefs held from different standpoints requires additional argument.

Whether it is plausible to think that two beliefs from different standpoints are candidates for conflict will depend on the particular criterion offered for believing something from each standpoint. Consider then the suggestion that different sets of interests distinguish the two types of standpoints and corresponding types of belief. It seems that the fact that we rely on one belief when we are engaged in one kind of task and we rely on its contrary when we are engaged in another (assuming that such a practice is possible) does not absolve us of irrationality. For example, if someone relies on his belief that God exists when engaged in prayer, and relies on his belief that God does not exist when deciding how to plan his future, I believe most would say that he is irrational. Thus, even if the "interests" criterion were to succeed in sorting beliefs appropriately, it lacks the resources to explain why it is not irrational to believe that we are free and not determined and to believe that we are determined.

---

[10] *The View from Nowhere* (New York: Oxford, 1986). According to Nagel, a standpoint or view is "more objective than another if it relies less on the specifics of the individual's makeup and position in the world, or on the character of the particular type of creature he is" (p. 5).

But perhaps some other criterion could be given which sorts beliefs appropriately within distinct standpoints, and, in so doing, absolves us from the charge of irrationality in believing the deliverances of reason. What could such a criterion be? The following passage by Korsgaard offers two suggestions: "Although the beliefs are theoretical in form—the will is free, there is a God—their basis and their function are practical."[11] Perhaps the two standpoints (and the two corresponding kinds of beliefs) are distinguished either by the way they are justified or by their function.

(B) *The justification criterion.* The idea that the two standpoints are distinguished by the types of justification for their corresponding beliefs has the apparent advantage of being able to provide a principled distinction between the belief that one is free and the belief that one is not free. For example, it might be argued that the belief that one is determined and not free is justified by reasoning about the possibility of experience, while the belief that one is not determined and free is justified in a practical way by its being required to achieve the ends associated with practical reasoning. There is reason to doubt, however, whether the "justification" criterion distinguishes between the beliefs in question in a way that will account for the apparent *irrationality* in believing them simultaneously.

Suppose that I believe *p* on the basis of being told that *p* by a reliable source and that I believe *not-p* on the basis of working through a proof. Despite the difference in the types of justification, I would surely be irrational in holding both beliefs simultaneously. Or, consider a different set of justification types which is very similar to the one described above in connection with the belief in freedom case. There is the sort of "practical" justification which applies to

---

[11] "Morality as Freedom," pp. 35-36.

Blaise Pascal's belief in the existence of God (because he stands to gain eternal happiness if he believes and his belief turns out to be true), or William James's belief that he can jump over a wide ravine (because he will have a better chance of succeeding if he believes than if he does not). We can imagine that each also has the contrary belief that is justified by theoretical or scientific reasons. In this case, too, a difference between two types of justification does not absolve either Pascal or James of irrationality. These are cases in which it may be "rational to be irrational," but they are not cases in which there is no irrationality.

It is possible to seek a more sophisticated categorization of justification types that applies in the belief in freedom case, but there is no particular reason to think that added sophistication will make a relevant difference here. Thus, providing a pair of justification types that succeed where others clearly fail is a challenge that must be met by the advocate of this two-standpoints view.

(C) *The function criterion.* Finally, consider the suggestion that a distinction between different points of view should be drawn by means of the functional role of their corresponding beliefs. Before we can assess this suggestion, clarification of the notion of "function" is required. In the sense of "function" according to which one intentionally employs something for a certain purpose, it is not clear that beliefs can have functions. It may be that we can have an understanding of what it would be to make an assumption for a certain purpose, but not of what it would be to believe for a certain purpose. Yet there is another sense of "function," namely, that according to which something produces a certain result via a normally functioning mechanism. In that sense, the idea that our beliefs have different functions is intelligible: some beliefs may be operative in some activities, other beliefs in others.

At first glance, this suggestion appears to face difficulties similar to those faced by the "interests" criterion. We can ask, for example, why it is that the theoretical belief that one is determined and not free should be used only for the purpose of explaining the world and not also for the purpose of deciding how to act. But one might defend the "function" criterion in the following way: the belief that one is not free simply *cannot* be used in order to decide how to act—one could not engage in that very activity if one employed that belief. Therefore, in deciding how to act, one employs only the belief that one is free and undetermined.

The function criterion still faces problems. Arguably, one employs practical reasoning in all sorts of contexts and for all sorts of purposes. If one employs the belief that one is free for every purpose, including that of determining, say, whether one *can* act in a certain way, then one will appear to be making use of the two contradictory beliefs in the very same activity. Of course, the contradictory beliefs will operate on two distinct levels: the belief that one is free allows one to engage in the very activity of deciding to deliberate about whether one can act in a certain way, while the belief that one is not free will play a role of supporting one's actual reasoning to the conclusion about whether one can act in a certain way. Nevertheless, there is a sense in which one "uses" both beliefs in engaging in a single activity.

It must be admitted that this argument raises many questions about the individuation of activities and the relation between practical and theoretical reasoning. Yet it is possible to set aside these interesting questions here. For even if the "function" criterion were to sort beliefs appropriately, the question remains whether it would succeed in absolving us of irrationality in holding the beliefs produced by reason. According to the "function" criterion, rational deliberators make use of the belief that they are free for certain purposes and can make use of the belief that they are not free for distinct purposes. This situation would seem to constitute a paradigm of irrationality. For example, if I believe that there is no external world while attempting to acquire a great benefit (for example, winning a million dollars), while at the same time I believe that there is an external world while engaged in obtaining an accurate picture of my situation, I am not fully rational. It may be that I am "practically" rational, in that I am attaining my ends in an efficient way, but this is not sufficient to show that my having contradictory beliefs does not constitute irrationality. In the case in which I hold contradictory beliefs that function in different ways, it seems that either one or both beliefs must fail to be justified, and hence that I am irrational in holding both of them. Thus, the function criterion fails to sort beliefs in a way that makes it rational to hold both beliefs produced by reason.

II. CONCLUSION

In sum, there are three problems for the two-standpoints account. The first is that the idea of believing from a standpoint may not be intelligible on reflection. The second is that it is not easy to identify a plausible criterion for deciding when a belief is held only from one standpoint. Thus, it is not clear why the belief that one is determined and not free is not held from the same standpoint as one's belief that one is free and not determined. In the absence of such a criterion, simply distinguishing between the two standpoints does not show that we can avoid irrationality in believing the results of reason. Finally, even if there were a criterion for deciding when a belief is held from only one standpoint (so that we could say that each of our beliefs is held exclusively from different standpoints), it would not follow that being in such a state is perfectly rational. Nagel's two-standpoints view provides a clear example of a view in which the appeal to different standpoints fails to protect against conflict. Further, it does not seem that any of the criteria we have considered fares better than Nagel's in dissolving the apparent conflict between the beliefs Kant takes to be the products of reason.

Now, it is possible that there is a version of the two-standpoints account other than the ones I have considered which adequately addresses these problems. But I do not believe that we have yet been presented with a good argument to the effect that beliefs from two different standpoints can fail to be candidates for conflict. Further, there is some reason to think that such an argument will not be forthcoming. For it is one thing to accept that there is no conflict between believing *p* and desiring *not-p*, but it is a very different thing to accept that there is no conflict between the deliberative belief that *p* and the scientific belief that *not-p*. Given that in each case we have an attitude of epistemic commitment—a commitment to the fact that the world is a certain way—it certainly seems irrational to have two such attitudes toward contradictory propositions.

DANA K. NELKIN

Florida State University

BOOK REVIEWS

First-order Logic: An Introduction. LEIGH S. CAUMAN. Hawthorne, NY: Walter de Gruyter, 1998. 343 p. Paper $29.95.

Leigh Cauman has written an elegant, no-frills, and no-nonsense logic text. As we would expect from the Managing Editor Emerita of this JOURNAL, the book is clean in style and editing. (I reacted with sinful glee when a student found an error on page 44: 'p→r', not 'p→q'.)

The book starts from the rudiments and ends with a lucid chapter on identity and definite descriptions. Cauman uses both natural deduction and truth trees. There is an appendix which, in the course of clarifying the notion of a variable, engages students with the philosophy of logic. The book covers what should be covered in a solid, one-semester introductory course in logic.

The text unfolds along traditional lines—propositional logic, monadic predicate logic, full (relational) first-order logic, the logic of identity. (Broadly, Cauman's model is the first edition of W. V. Quine's *Methods of Logic*,[1] though her text is less advanced.) Each chapter ends with a set of reminders and exercises (with an instructor's manual).

Although the work is mainly along standard lines, here and there our attention is arrested by thoughtful deviations. For example, the second rule she introduces is conditional proof. Despite its complexities and dangers, which is presumably why it is typically presented late, conditional proof is one of our most basic forms of reasoning—we suppose a proposition to discover what consequences follow. By introducing conditional proof early, Cauman draws upon the student's own reasoning, and so can build upon it. Here, as elsewhere, Cauman's innovations struck me as helpful and insightful.

I found especially valuable her regular explanations of proofs by talking them through. Here is an example where Cauman makes sense of why a certain step in a proof is fallacious. The argument is:

> Each of us loves someone or other.
> Therefore, there is someone whom everybody loves.

After setting out the fallacious formal 'proof', she talks it through:

> * 1 Everyone loves someone or other. PREM

---

[1] New York: Holt, 1950; Cambridge: Harvard, 1982, 4th edition.

0022-362X/00/9710/577–80

* 2 So Albert, for example, loves someone or other. 1 UI
* 3 Let's call her Beatrice; Albert loves Beatrice. 2 EI (b)
* 4 But Albert (see step 2) was just anybody; so, since Albert loves Beatrice, everybody loves Beatrice. 3 UG (a)
* 5 There is someone (Beatrice) whom everybody loves. 4 EG (238)

Students come to appreciate the fallacy at step 4—and thus the restrictions on UG—through assimilating it to an error that they already recognize.

When a hurdle is encountered in either ordinary reasoning or formalization or proofs or truth trees, Cauman turns to a different one of these to aid in surmounting the hurdle. So students will construe the introduction of each new method not as a further burden on them, but as a further resource. For example, she observes that we sometimes read a universal claim as having existential import because we overlook the conditional nature of an assertion of existence. She proceeds to provide two translations for:

> If someone telephoned it was John (317).

one via a universal quantifier and the other via a conditional (with a small-scope existential quantifier). Finally, she recalls students to the tree method to demonstrate the claimed equivalence.

More than many texts, Cauman's work is governed by the goal of improving ordinary reasoning. The first paragraph of her Introduction ends:

> ...there is a practical as well as theoretical reason for the study of logic: to train the mind (1).

Because this worthy aim pervades the text, and the text prods us to think carefully about pedagogy, I want to pursue briefly the prospects for logic as a device to improve reasoning.

A simple, direct way that teaching logic holds out promise to improve reasoning would be if the structure of ordinary reasoning is in good part that of (first-order) formal logic. But the promise of this simple, direct approach is cast in doubt by a number of considerations. First, to reconstruct and test realistically complex arguments via formal logic requires an unusual level of achievement and an unfeasible amount of time to engage in the task. Second, Gilbert Harman[2] argues that logic has no "special relevance to reasoning" since logic is the study of implication relations, while reasoning and inference are concerned with change in belief under practical constraints. Third, psychologists have exposed errors or fallacies in everyday deductive reasoning.[3]

---

[2] *Change of View* (Cambridge: MIT, 1986).

Now this third consideration, which is the one I focus on, might seem to supply a strong reason to study logic. But the psychologists do not find logic courses that helpful in eliminating the errors or fallacies, even when they do propose various ways to improve reasoning.

The most famous of these studies (the "Wason selection-task" or "four-card problem") shows shortcomings in conditional reasoning. For your pleasure, let me present a recent, and much less familiar or studied, example due to Philip Johnson-Laird.[4] Does the following premise:

> If Cauman is right, learning logic improves reasoning; *or else* if Cauman is not right, learning logic improves reasoning.

imply that:

> learning logic improves reasoning?

Take a minute. I'll wait.... The answer is "no." In fact, the conclusion is incompatible with the premise. A large number of you got it wrong (or would have except for my cuing you to think twice). Although this example seems more a mere verbal trick or brain teaser than the selection task, there is a link of wide scope: we tend not to generate (or maintain) explicit representations of negative implications.[5]

A number of times I have presented the selection task on exams in my logic courses, along with the following kinds of true-false problems:

> If the premises of the argument are inconsistent, the argument must be valid.
> If an argument is deductively valid and we add a new premise, the resulting argument may be invalid.
> If '$p \rightarrow q$' is true, then the argument from 'p' to 'q' is valid.

The results are better for these true-false questions than the dismal scores on the selection task. Nevertheless, students still do considerably worse on the former than on the more standard problems. These true-false questions test students' command of logical concepts. The mediocre results accord with findings that students in any discipline can do well on exams that stick closely to those specifically taught in class or text, but that their grades sink when problems are varied.

---

[3] Some researchers hold that hardly any of our actual reasoning is deductive, rather than inductive or inference to the best explanation or (nonmonotonic) default reasoning.

[4] For this example, and a grip on an important facet of the discussion in psychology, see the exchange between Johnson-Laird and Lance Rips beginning with Johnson-Laird's: "Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*," *Minds and Machines*, VII (1997): 387-407, 409-24, 425-32.

[5] The currently preferred form of the problem is to drop the 'or else' and just introduce the conditionals with:

Only one of the following assertions is true:

...

The same results are found. Expressing the problem in the form '$p \rightarrow q$ or not-$p \rightarrow q$, but not both' has not been tried, and it is thought to introduce linguistic confusions (personal communication with Johnson-Laird).

Although this is a common finding, we ought to take little comfort in the shared misery, given the justification of logic instruction under consideration. For that justification is to develop reasoning, which requires resilience under a good deal of alteration in content.

Assuming that my failure indicates a shortcoming of a standard logic course, what ought we to conclude about the aspirations of such a course? The results do not undermine the justification for teaching logic either as a requirement for philosophy majors, since that is a prerequisite to advanced work, or as a discipline of knowledge within the liberal arts, or as providing a useful or necessary skill for other areas of study (especially computer science). But only its value toward improving thinking—to help us, as Cauman concludes, "to think straight" (335)—would justify a recommendation of the study of logic to many students.

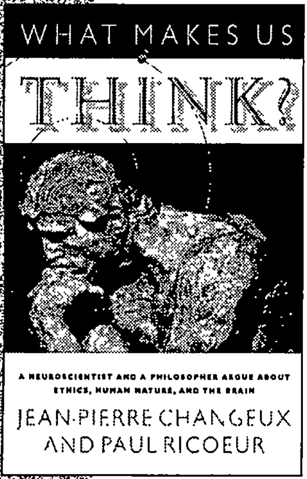
We should neither take for granted the value of logic in improving everyday reasoning nor dismiss it as empirically refuted. Logic teaches a regulative ideal for arguments in which premises and assumptions are broken down into constituents, expressed explicitly, and each inference is justified. Salience is brought to a number of simple valid patterns—modus ponens, conditional proof, modus tollens, disjunctive syllogism, reductio ad absurdum—which dominate inference, and thus provide structures for the generation and analysis of arguments. Central critical concepts are clarified and brought on-line: consistency, validity, implication. Fallacies are susceptible to sharp representation. Students can be brought to recognize how natural assumptions of reasoning and conversation facilitate fallacious thinking—for example, scope fallacies of negation, quantification, and modality.

I take these potential benefits—among others—for ordinary reasoning to be sustainable in the face of the empirical evidence. Nevertheless, that evidence should guide us in the improvement of logic courses by using more realistic examples (for example, translation/proof problems with irrelevancies) and by placing more emphasis on the workings of counterexamples. And we should seek texts with the pedagogical sensitivity and passion for good thinking which are exhibited in Cauman's *First-Order Logic*.

JONATHAN E. ADLER

Brooklyn College and the Graduate School/CUNY

New from Princeton


WHAT MAKES US
THINK?
A NEUROSCIENTIST AND A PHILOSOPHER ARGUE ABOUT ETHICS, HUMAN NATURE, AND THE BRAIN
JEAN-PIERRE CHANGEUX
AND PAUL RICOEUR

What Makes Us Think?

A Neuroscientist and a Philosopher Argue about Ethics, Human Nature, and the Brain

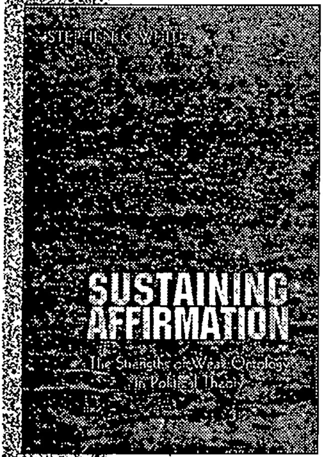
Jean-Pierre Changeux
and Paul Ricoeur

Translated by M. B. DeBevoise

Will understanding our brains help us to know our minds? Or is there an unbridgeable distance between the work of neuroscience and the workings of human consciousness? In a remarkable exchange between neuroscientist Jean-Pierre Changeux and philosopher Paul Ricoeur, this book explores these questions.

"These two amazing minds at work make for a fascinating look at the who, what, and how thought happens."
—*Choice*

Cloth $29.95 ISBN 0-691-00940-6



Sustaining Affirmation

The Strengths of Weak Ontology in Political Theory

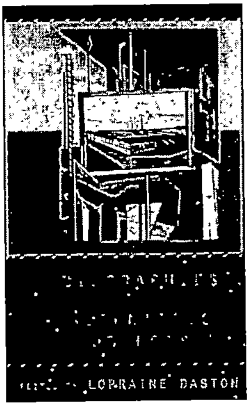
Stephen K. White

In light of many recent critiques of Western modernity and its conceptual foundations, the problem of adequately justifying our most basic moral and political values looms large. Without recourse to traditional ontological or metaphysical foundations, how can one affirm—or sustain—a commitment to fundamentals? The answer, according to Stephen White, lies in a turn to "weak" ontology.

Paper $15.95 ISBN 0-691-05033-3
Cloth $49.50 ISBN 0-691-05032-5

Princeton University Press
800-777-4726 • WWW.PUP.PRINCETON.EDU

the life of science

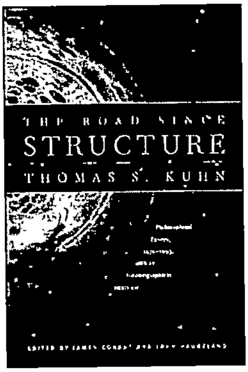


Biographies of Scientific Objects

Edited by Lorraine Daston

"Why do scientists select or create some things but not others to be the objects of their investigations? This impressive and highly original collection of essays, ranging from sociology and anthropology to theoretical physics and mathematics, looks at the varied and often surprising processes—intellectual, practical, and social—through which the subject matters of the sciences are brought into being, and, on occasion, may pass away. Biographies of Scientific Objects is a tour de force."—Nicholas Jardine, University of Cambridge

Paper $19.00



The Road since Structure

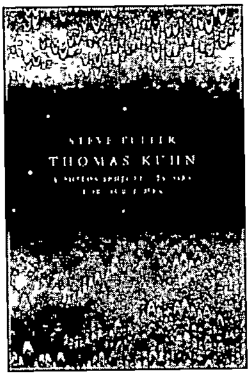
Philosophical Essays, 1970–1993, with an Autobiographical Interview

Thomas S. Kuhn

Edited by James Conant and John Haugeland

This volume is the fullest record we now have of the new direction Thomas Kuhn was taking during the last two decades of his life. The first part of the book consists of essays in which Kuhn refines the basic concepts set forth in *The Structure of Scientific Revolutions*, and in part two, Kuhn replies to many of the criticisms of his earlier work. The third part is the transcript of a remarkable and singular interview that the normally reticent Kuhn conducted just before his death.

Cloth $25.00



Thomas Kuhn

A Philosophical History for Our Times

Steve Fuller

"The resonance of the phrase 'paradigm shift' amply testifies to the profound influence of Thomas Kuhn's book *The Structure of Scientific Revolutions* (1962). For Fuller, that influence cannot end soon enough. . . . A brilliant analysis deserving as wide a readership as the acclaimed book it critiques."—Booklist, starred review

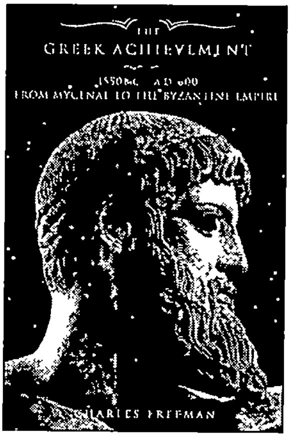
Cloth $35.00

The University of Chicago Press 5801 South Ellis Avenue, Chicago, IL 60637
www.press.uchicago.edu

CHARLES FREEMAN

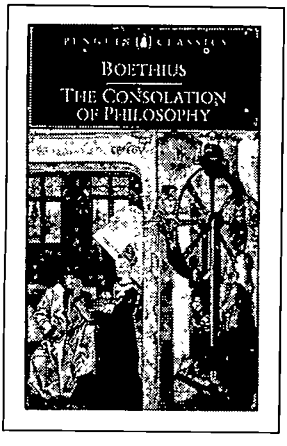
"An ambitious work...commendably up to the minute."—Paul Cartledge, *History Today.* "A generous account, graceful yet panoramic."—Robert Fagles. Freeman paints a sweeping panorama of the ancient Greeks' world and provides a rich, contemporary overview of their enduring contribution to world civilization. Maps and illustrations throughout, 16 pp. b/w photographs, 8 pp. color photographs.

Penguin 512 pp. 0-14-029323-X $20.00



THE GREEK ACHIEVEMENT
THE FOUNDATION OF THE WESTERN WORLD

THE CONSOLATION OF PHILOSOPHY



BOETHIUS

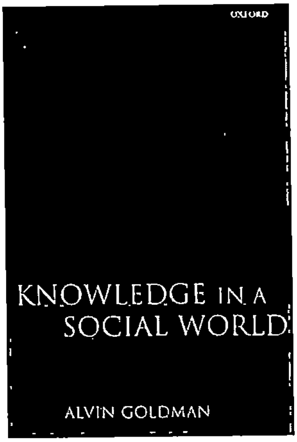
REVISED EDITION, Translated and with an Introduction and new Preface by Victor Watts. The clarity of Boethius's thought and his breadth of wisdom made this work—a dialogue of alternating prose and verse between an ailing prisoner and his "nurse," Philosophy—hugely popular throughout medieval Europe; his ideas suffused the thought of Chaucer and Dante. Watts's translation makes this text accessible to the modern reader while losing nothing of Boethius's poetic artistry and philosophical brilliance.

Penguin Classic 192 pp. 0-14-044780-6 $13.95

PENGUIN PUTNAM INC.
ACADEMIC MARKETING DEPT. • 375 HUDSON STREET • NY, NY 10014-3657 • www.penguinputnam.com

KNOWLEDGE IN A SOCIAL WORLD

ALVIN I. GOLDMAN



"An important work of monumental scope. Its central concern is the ways in which social practices and familiar institutions contribute to and undermine the pursuit of knowledge.... A book of singular importance to lawyers, political theorists, social and natural scientists as well as to educators and theorists of education. It is a major contribution to all these fields and not just because of its insights into them, but for its accessibility to intelligent practitioners as well. It may well be the most significant interdisciplinary philosophy book of the decade."

—Jules Coleman, *Yale Law School.*

"Alvin Goldman, the premier epistemologist of the past two decades, has written a pioneering book that will define the field of social epistemology. Scholars will learn from his judicious and lucid proposals, and they will be wrestling for years with the exciting and important problems he raises."

—Philip Kitcher.

"...belongs to a sparse tradition in epistemology given to the epistemic evaluation of social institutions. Landmarks in this tradition include Bacon's *The Advancement of Learning*, Locke's *Some Thoughts Concerning Education*, and Dewey's *Social Inquiry*. Goldman's book is the equal of these influential volumes in breadth of vision, mastery of relevant contemporary thought, and practical utility, and surpasses them in theoretical depth. It is a work crafted with supreme philosophical skill."

—Frederick Schmitt.

Knowledge in a Social World offers a philosophy for the information age. Alvin Goldman explores new frontiers by creating a thoroughgoing *social* epistemology, moving beyond the traditional focus on solitary knowers. Against the tides of postmodernism and social constructionism Goldman defends the integrity of truth and shows how to promote it by well-designed forms of social interaction. From science to education, from law to democracy, he shows why and how public institutions should seek knowledge-enhancing practices. The result is a bold, timely, and systematic treatment of the philosophical foundations of an information society.

1999 424 pp.; 3 line illus paper $19.95 cloth $70.00

OXFORD UNIVERSITY PRESS

OXFORD UNIVERSITY PRESS

KANT'S THEORY OF MIND

An Analysis of the Paralogisms of Pure Reason

Second Edition

KARL AMERIKS

This seminal contribution to Kant studies, originally published in 1982, has now been brought up to date with a substantial preface and postscript.

2000 392 pp. cloth $65.00 paper $19.95

CULTURAL OTHERNESS

Correspondence with Richard Rorty

Second Edition

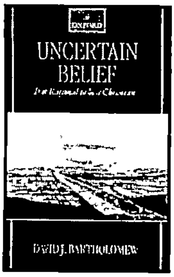
ANINDITA NIYOGI BALSLEV

"The letters...are a fine example of what they advocate—genuine dialogue between real interlocutors, and not fantasies of what some 'other' might say."
—Cleo McNelly Kearns, Rutgers University.

(AAR Cultural Criticism Series 4)

(An American Academy of Religion Book)

2000 128 pp. paper $16.95



New in paperback!

UNCERTAIN BELIEF

Is it Rational to be a Christian?

DAVID J. BARTHOLOMEW

"Recommended for all college collections."—*Choice*. Arguing from a probabilistic perspective, Bartholomew provides an answer to the fundamental question: is it rational to be a Christian?

1996 (paper 2000) 304 pp.
paper $19.95 cloth $35.00

BERNARD OF CLAIRVAUX

G. R. EVANS

In this book the renowned medievalist G.R. Evans provides a concise introduction to St. Bernard of Clairvaux (1090-1153). The only available introduction to Bernard's life and thought, this latest addition to the *Great Medieval Thinkers* series will appeal to a wide audience of students and scholars of history and theology.

(Great Medieval Thinkers)

2000 232 pp. paper $18.95 cloth $45.00

New in paperback!

VIRTUAL REALISM

MICHAEL HEIM

"Heim's views on technology are far removed from the routine rhetoric of cyberhype. With one foot in the book-crammed world of the humanist scholar and the other in the factories of the future, his ideas are always sensitive to both the euphoria and the anxiety which new technologies engender. Above all, his approach is pragmatic: accepting our technological fate and examining different ways to absorb its impact."—Mark Nixon, *Twenty First Century*.

1998 (paper 2000) 264 pp.; 22 4-color illus, 16 b/w illus paper $16.95 cloth $26.00

New in paperback!

THE SCENES OF INQUIRY

On the Reality of Questions in the Sciences

NICHOLAS JARDINE

"Offers many refreshing and sensible insights to current debates about truth, scientific realism, and particularly the historical evaluation of scientific progress."
—*Philosophical Review*.

1991 (paper 2000) 328 pp. paper $24.95

INDETERMINATE IDENTITY

Metaphysics and Semantics

TERENCE PARSONS

Parsons presents a lively and controversial study of philosophical questions about identity. He claims that there is genuine indeterminacy of identity in the world, articulating such a view in detail and defending it from a host of criticisms.

2000 210 pp.; 26 figs. $49.95

The final volume of Ward's acclaimed tetralogy!

RELIGION AND COMMUNITY

KEITH WARD

This book explores the relationship between religion and society, and discusses the ways in which the major world religions need to adapt to the modern world. Keith Ward looks at different forms of religious community, then proposes a radical vision of the church as a person-affirming, world-transforming society within the emerging global community.

2000 376 pp. paper $19.95 cloth $78.00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, 1-800-387-8020. www.oup.com

Critical Inquiry

Editor
W. J. T. Mitchell

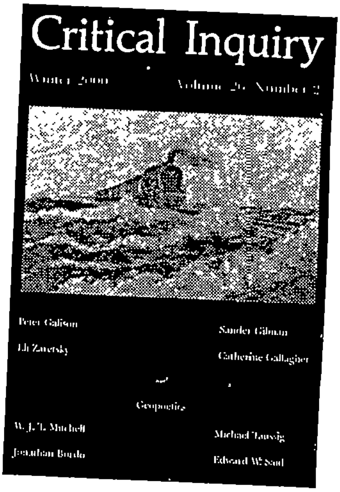
An intimate exploration of criticism...

A challenging look ...

Founded 26 years ago, *Critical Inquiry* is an interdisciplinary journal devoted to publishing the best critical thought in the arts and humanities. Combining a commitment to rigorous scholarship with a vital concern for dialogue and debate, the journal presents articles by eminent critics, scholars, and artists on a wide variety of issues central to contemporary criticism and culture.

At scholarly criticism and culture ...

In *CI* new ideas and reconsideration of those traditional in criticism and culture are granted voice. International authors express inspiring and often provocative ideas on hegemony and variant cultural styles in literature, art, architecture, film, history, music, philosophy, science, and more. Counterparts debate definitions and reflections of the most brilliant, infamous, and popular moments, past and present. In *CI*, authors entertain and challenge while extolling on such issues as Einstein's philosophy of time; early modern women's writing; political-religious identity in the Middle East; and other rigorous subjects. *CI* comes full circle with the electrically charged debates between contributors and their critics.



Order securely through the *CI* Web site
www.journals.uchicago.edu/CI

Critical Inquiry (ISSN 0093-1896) is published quarterly. Regular one-year subscription rates are $37 individuals, $25 students, $100 institutions. Single copies $8 individuals, $25 institutions. Outside the US, please add $5 postage. Canadian residents please add 7% GST and $5 postage. MasterCard and Visa payments accepted. Credit card customers may fax orders to (773) 753-0811. To order your subscription send check, purchase order, or credit card type, account #, exp. date, telephone #, and signature to the University of Chicago Press, Journals Division, Department SS0SA, P.O. Box 37005, Chicago, IL 60637 USA. For more information, please visit our web site at www.journals.uchicago.edu/CI.

The University of Chicago Press
Journals Division, Box 37005, Chicago, IL 60637

6/00

the review of

metaphysics

a philosophical quarterly

MARCH 1998 | VOL. LI, No. 3 | ISSUE No. 203 | $12.00

articles

ROBERT SOKOLOWSKI — *The Method of Philosophy: Making Distinctions*

JOHN F. WIPPEL — *Thomas Aquinas and the Axiom that Unreceived Act is Unlimited*

NORMAN J. WELLS — *Descartes and Suárez: A Tale of Two Readings*

JEFFREY E. BROWER — *Abelard's Theory of Relations: Reductionism and the Aristotelian Tradition*

JACOB HOWLAND — *The Republic's Third Wave and the Paradox of Political Philosophy*

ALBERT E. GUNN AND STAFF — *Summaries and Comments*

book reviews *philosophical abstracts* *announcements*

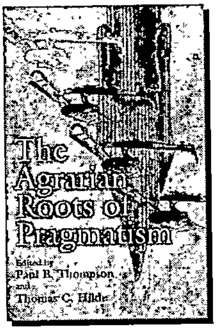
Individual Subscriptions $28.00 Institutional Subscriptions $45.00 Student/Retired Subscriptions $18.00

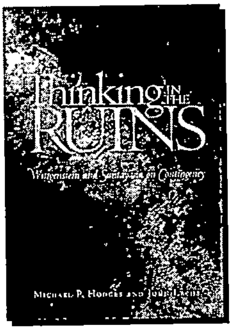
VISA AND MASTERCARD ACCEPTED

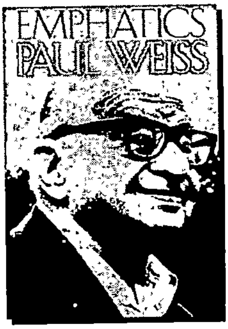
Jude P. Dougherty, Editor
The Catholic University of America, Washington D.C. 20064
email: revmeta@nlcom.com

The Vanderbilt Library of American Philosophy

Recently Published


The Agrarian Roots of Pragmatism
Edited by Paul B. Thompson and Thomas C. Hilde


Thinking in the Ruins
Wittgenstein and Santayana on Contingency
Michael P. Hodges and John Lachs



The Agrarian Roots of Pragmatism
Edited by
Paul B. Thompson and
Thomas C. Hilde

"*The Agrarian Roots of Pragmatism* represents a unique contribution to the history of American thought, one that has been entirely too neglected. This collection promises to be an especially important resource for scholars of rural life and for scholars and activists involved in the reform of agriculture."
—RICHARD HAYNES

352 pages
Cloth $39.95

Thinking in the Ruins
Wittgenstein and Santayana on Contingency
Michael P. Hodges
and John Lachs

"This book offers a stimulating and insightful assessment of the philosophies of Santayana and Wittgenstein in coordinative apposition. By interactively juxtaposing their ideas the authors are able to throw interesting new light upon the views of these two giants of twentieth-century philosophy."
—NICHOLAS RESCHER

144 pages
Cloth $29.95

Emphatics
Paul Weiss

"Metaphysics in the grand style."
—LIBRARY JOURNAL

"New ideas are rare in philosophy. But when they occur, are imagined, articulated, and defended with systematic intelligence, they make us see everything differently. Weiss's notion of 'emphatics' is such a rare new idea, and it gives us a new system of philosophy from this great master of systematic philosophy."
—ROBERT C NEVILLE

272 pages
Cloth $34.95

Vanderbilt
1-800-627-7377
Vanderbilt University Press • www.vanderbilt.edu/vupress

Best in Philosophy from Cambridge

Fichte: *Foundations of Natural Right*
Johann Gottlieb Fichte
Cambridge Texts in the History of Philosophy
0-521-57301-7 Hardback $64.95
0-521-57591-5 Paperback $22.95

Kant and Modern Political Philosophy
Katrin Flikschuh
0-521-66237-0 Hardback $54.95

Making Moral Sense
Beyond Habermas and Gauthier
Logi Gunnarsson
Cambridge Studies in Philosophy
0-521-78023-3 Hardback $59.95

Coleridge, Philosophy and Religion
Aids to Reflection and the Mirror of the Spirit
Douglas Hedley
0-521-77035-1 Hardback $64.95

Probability, Econometrics and Truth
The Methodology of Econometrics
Hugo Keuzenkamp
0-521-55359-8 Hardback $69.95

Heidegger, Language and World-Disclosure
Cristina Lafont
Modern European Philosophy
0-521-66247-8 Hardback $59.95

Objectivity in Law and Morals
Brian Leiter, *Editor*
Cambridge Studies in Philosophy and Law
0-521-55430-6 Hardback $64.95

The Cambridge Translations of Medieval Philosophical Texts
Volume 2: Ethics and Political Philosophy
A.S. McGrade, John Kilcullen, *and* ***Matthew Kempshall,*** *Editors*
0-521-23625-8 Hardback $85.00
0-521-28082-6 Paperback $29.95

Aristotle's Theory of Language and Meaning
Deborah K.W. Modrak
0-521-77266-4 Hardback $59.95

The Cambridge Companion to Malebranche
Steven Nadler, *Editor*
Cambridge Companions to Philosophy
0-521-62212-3 Hardback $54.95
0-521-62729-X Paperback $19.95

Bounds of Justice
Onora O'Neill
0-521-44232-X Hardback $54.95
0-521-44744-5 Paperback $19.95

Plato: *The Republic*
G.R.F. Ferrari, *Editor*
Cambridge Texts in the History of Political Thought
0-521-48173-2 Hardback $37.95
0-521-48443-X Paperback $10.95

Voltaire: *Treatise on Tolerance*
And Other Writings
Simon Harvey, *Editor*
Cambridge Texts in the History of Philosophy
0-521-64017-2 Hardback $54.95
0-521-64969-2 Paperback $19.95

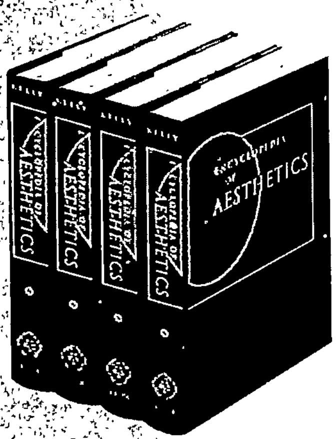
Available in bookstores or from
CAMBRIDGE
UNIVERSITY PRESS

40 West 20th Street, N.Y., NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cup.org
AmEx/MasterCard/VISA accepted.
Prices subject to change.

"An immediate reference classic."*

ENCYCLOPEDIA OF AESTHETICS

Editor in Chief: Michael Kelly



- 600 original articles
- 450 distinguished contributors
- 100 illustrations
- Bibliographies, cross-references, and an exhaustive index

"*A unique work of splendid scholarship . . . highly recommended for any library serving a liberal arts program.*"—Choice

"*It is a magnificent achievement and deserves a very wide readership.*" —Journal of Philosophy

"*Both authoritative and accessible...an illuminating guide to historical and contemporary scholarship.*"—Library Journal

In four volumes and 600 original articles, this landmark reference source surveys an extraordinary range of ideas about the arts in human life.

From Theodor Adorno to Frank Lloyd Wright, African aesthetics to video, the articles in the Encyclopedia examine a remarkable range of ideas and issues:

- the full breadth of critical thought on the arts—from classical philosophy to current literary theory
- in-depth survey of Western and non-Western traditions
- the political, social, economic, and ideological contexts of creativity
- the critical vocabularies of many forms, genres, and periods, from abstract expressionism to jazz to video
- the work of dozens of leading thinkers

2,240 pages • Four volumes
0-19-511307-1 • $495.00

OXFORD UNIVERSITY PRESS
198 Madison Avenue, New York, NY 10016
To order 1.800.451.7556 • www.oup-usa.org

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NUMBER 11
NOVEMBER 2000

page		
581	*The Social Character of Testimonial Knowledge*	Paul Faulkner
602	*Reliabilism Leveled*	Jonathan Vogel
624	*A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem*	Juliet Floyd and Hilary Putnam
633	NEW BOOKS: PAPERBACK REPRINTS	

UNIVERSITY LIBRARY
ALIPUR
CAMPUS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCVII, NO. 11, NOVEMBER 2000

THE SOCIAL CHARACTER OF TESTIMONIAL KNOWLEDGE*

Through communication, we form beliefs about the world, its history, others, and ourselves. A vast proportion of these beliefs we count as knowledge. We seem to possess this knowledge *only* because it has been communicated. If those justifications which depended on communication were outlawed, all that would remain would be a body of ill-supported prejudice. The recognition of our ineradicable dependence on testimony for much of what we take ourselves to know has suggested to many that an epistemological account of testimony should be essentially similar to accounts of perception and memory. This is the premise I want to dispute.

Most would deny that perceptual knowledge is mediated. The acquisition of knowledge through communication is mediated in one obvious sense. In a case of successful perception, one knows that the world is a certain way because one is aware of the world being that way. But if we come to know that the world is a certain way through communication, we have no comparable awareness of how the world is. We know that the world is a certain way only because another has represented the world as being that way. As such, our acquisition of knowledge can depend on why the speaker represented the world to be that way and whether or not the speaker knows that the world is how he represented it to be. Thus, testimony is mediated in the sense that the intentions of another, and the justification possessed by another *are relevant to* the audience's acquisition of knowledge. These considerations have no parallel in either perception or memory.

---

* I am indebted to M.G.F. Martin and Matt Soteriou. My thanks to Matt Nudds and Tom Crowther.

0022-362X/00/9711/581–601

Nonetheless, there seems to be a growing consensus in philosophy that epistemological accounts of testimony, perception, and memory should be similar. John McDowell[1] writes:

> Just as one can capture a knower's justification for believing what he does by saying that he sees that things are thus and so, or that he remembers that things are thus and so, so one can capture a knower's justification—his knowledge-constituting standing in the space of reasons—by saying that he has heard from so-and-so that things are thus and so (*ibid.*, p. 210).

McDowell proceeds to give a disjunctive account of testimony, which accords, in all its essential aspects, with the disjunctive account he provides of perceptual knowledge.[2] Similarly, when Tyler Burge[3] considers our *entitlement* to rely on testimony, it turns out that:

> Many of the differences between content passing between minds and content processed by a single mind derive from differences in modes of acquisition and in necessary background conditions, that do not enter in the justification force underwriting entitlement (*ibid.*, p. 474).

This distinction—between a belief's genesis and its justification—allows Burge to drain the distinctive features of testimony of epistemic significance. Reliance upon testimony, as with reliance on memory or perception, is simply reliance upon a "resource for reason."[4]

I think that this approach to testimony gets its epistemology drastically wrong. My aim here is to provide an epistemological theory of testimony that recognizes its difference from perception and memory. I take as my starting point the work of Burge, focusing on his theory of testimony for a number of reasons. It provides an extreme case of aligning testimony with perception and memory; it is well argued; and unlike McDowell's theory, it lies within the mainstream of epistemology.

My argument here is developed in five sections and a conclusion. In section I, I consider Burge's *acceptance principle*, which states our entitlement to accept testimony credulously. The way in which Burge introduces the acceptance principle presupposes that, in certain respects, testimony is similar to perception and memory. In section II, I question this presupposition and consider a way in which testimony is quite distinctive. This motivates a converse principle stating that we must support our acceptance of testimony with reasons. I name this the *principle of assent.* In section III, I argue that the principle of assent is consistent with the justification of the acceptance principle and that it is compatible with a particular interpretation of the acceptance principle. I name this *re-interpreted* acceptance principle the *principle of warrant.* In section IV, I combine the principles of warrant and assent to form a *hybrid theory of testimony* and outline how this theory relates to existing epistemological accounts of testimony. In section V, I consider two different ways of understanding the principle of warrant. These provide two models of the *social nature* of testimonial justification. Finally, I conclude by considering the dynamic that exists between individual knowers and communities.

---

[1] "Knowledge by Hearsay," in Bimal Krishna Matilal and Arindam Chakrabarti, eds., *Knowing from Words* (Boston: Kluwer, 1994), pp. 195-224.

[2] "Criteria, Defeasibility and Knowledge," *Proceedings of the British Academy*, LXVIII (1982): 455-79.

[3] "Content Preservation," *Philosophical Review*, CII, 4 (1993): 457-88.

[4] Michael Dummett stresses the similarity of testimony and memory; see "Testimony and Memory" in his *The Seas of Language* (New York: Oxford, 1993), pp. 411-28. For an emphasis on the general epistemic similarity of these faculties see C. A. J. Coady, *Testimony: A Philosophical Study* (New York: Oxford, 1992).

I. BURGE'S ACCEPTANCE PRINCIPLE

On Burge's view, a belief is justified by the reasons that support it. He distinguishes two species of justification: argumentative justifications[5] and entitlements. Argumentative justifications are arguments articulated by a believer to the truth of a proposition. The reasons articulated, therefore, "must be available in the cognitive repertoire of the subject" (*op. cit.*, p. 459). Entitlements state our epistemic rights to rely on certain cognitive practices. They are equally constituted by argument. But the articulation of such an argument requires a "God's eye" view of the cognitive practice.[6] Thus, entitlement is the "externalist analog of the internalist notion, justification" (*ibid.*, p. 3). Entitlements "need not be understood by or even accessible to the subject."[7] This may best be illustrated by turning to the relevant entitlement: our entitlement to accept testimony credulously.

We are entitled to trust testimony. When a speaker presents his testimony as true, the audience is entitled to accept it, other things being equal. *A person is entitled to accept as true something that is presented as true and that is intelligible to him, unless there are stronger reasons not to do so* (*ibid.*, p. 467). This is Burge's *acceptance principle.* Unless the audience possesses reasons to reject something that is intelligibly presented as true, he is entitled to accept it.[8]

---

[5] The term 'argumentative justification' is Ernest Sosa's; see his " 'Circular' Coherence and 'Absurd' Foundations," in Ernest LePore, ed., *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (New York: Blackwell, 1986), pp. 387-97.

[6] Burge, "Computer Proof, A Priori Knowledge, and Other Minds," *Philosophical Perspectives,* XII, *Language, Mind and Ontology* (1998): 1-37, see p. 28.

[7] "Content Preservation," p. 458.

The justification that Burge offers for the acceptance principle is extremely abstract. Its starting point is that seeming intelligibility allows an audience to presume that a speaker is rational. Its crux is the claim that because a speaker is rational, an audience can presume that what the speaker presents as true is true. Its "general form" is that an audience is entitled to accept testimony because *"it is prima facie preserved (received) from a rational source, or resource for reason; reliance on rational sources—or resources for reason—is, other things being equal, necessary to the function of reason."*[9]

An immediate and difficult problem for this argument is that it can be rational to lie. As such, it seems that the (prima facie) *rational* characteristics of the source do not suffice to establish the (prima facie) *truth* of the source's presentations. That is, the crux of the argument is far from obviously true. Why should a speaker's rationality establish the presumption that, when his testimony is presented-as-true, it is true? Burge provides the following argument.

> One of reason's primary functions is that of presenting truth, independently of special personal interests. Lying is sometimes rational in the sense that it is in the liar's best interests. But lying occasions a disunity among the functions of reason. It conflicts with one's reason's transpersonal function of presenting the truth, independently of special personal interests. [Therefore]... Unless there is a reason to think that a rational source is rationally disunified—in the sense that individual interest is occasioning conflict with the transpersonal function of reason—one is rationally entitled to abstract from individual interest in receiving something presented as true by such a source (*ibid.*, p. 475).

One should be clear that insofar as a lie is rational, its rationality is *independent* of the liar's personal interest in that *any reasonable person* with such beliefs and desires would have acted accordingly. Burge's point, however, is that to understand a lie as rational, one needs to consider the liar's perspective. To understand the presentation of a truth as rational no such relativization is needed. A primary function of reason is to present truth and truth is not relativized to a believer or time. Thus, no relativization is needed because one can take, so to speak, "God's view." Let me call this the *transcendental perspective*: from this perspective, the presentation of truth is a primary function of reason.

---

[8] That is, the "entitlement holds unless there are stronger reasons (available to the person) that override it"—Burge, "Interlocution, Perception, and Memory," *Philosophical Studies*, LXXXVI (1997): 21-47, p. 22, note 4.

[9] "Content Preservation," p. 469.

Burge takes the transcendental perspective because he is concerned, not with our justification in acquiring testimonial beliefs, but with our entitlement to such beliefs. His conception of entitlements is a conception of a form of epistemic justification which articulates our right to rely on those cognitive practices which accord with the norms of reason. A focus on the norms of reason elides many of the differences between reliance upon perception, memory, and testimony. Thus the general form of the justification of the acceptance principle involves the equation of "rational sources" and "resources for reason."

Lying effects "a disunity among the functions of reason." The presentations of a rational source are not merely rationally supported—lies may be rationally supported. The presentations of a rational source may be presumed to be the presentations of reason. This is a strong presumption of rationality. This presumption underlies the acceptance principle and its justification. Burge states: "a *presupposition* [*not* a consequence] of the Acceptance Principle is that one is entitled not to bring one's source's sincerity or justification into question, in the absence of reasons to the contrary" (*ibid.*, p. 468; my emphasis).

II. THE DISTINCTIVENESS OF TESTIMONY.

The general form of the acceptance principle and its justification "are formulated so as to be neutral on whether what is 'presented as true' comes from another person" (*ibid.*, p. 474). If this is the case, I would argue, then the acceptance principle cannot apply to testimony. The fact that what is presented-as-true does come from another person is a distinguishing feature of testimony. I shall now argue that this feature renders testimony epistemologically distinctive.

In order to recognize the distinctiveness of testimony, one should start with the obvious point that we acquire testimonial knowledge through *communication.* Our purposes in communicating, however, are *not* merely to inform our audiences. Informing our audiences may be a secondary or derivative purpose—as when we tell someone something in order to get them to act—or it may be no purpose at all. Our purposes in communicating are many and varied. Few of these communicative purposes seem to *require* that a speaker believe what he expresses. For example, if I seek to reassure you, there seems to be no *need* for me to express something I believe; my concern is to express something you should find reassuring to believe and believable.

To a certain extent, noninformative and nondeceptive communicative purposes can be bracketed by speaking of presentations-as-true. Even if "testimony" and "presentation-as-true" are taken to be synonyms, however, story telling and jokes are still testimony. The existence of such speech acts makes it wrong to treat insincerity as the defeating condition in the acquisition of testimonial knowledge. One does not ordinarily acquire testimonial knowledge from a speaker who presents that *p* as true and yet believes it to be false. Yet if the speaker were merely fooling, it would be wrong to call him insincere. An insincere reassurance may be a lie, a joke is not. Suppose that a speaker presents that *p* as true. Let me call this the *speaker's testimony*. Let me say that, if the speaker believes that *p* is true and he is correct, it is the case that *p*, then the speaker is *dependable*.[10] Whilst, if the speaker believes that *p* is false, let me say the speaker is *artful*. Insincerity implies but is not implied by artfulness. Artfulness implies but is not implied by undependability: a speaker may be quite sincere and yet express a falsehood. Considerations of artfulness and dependability are relevant to the successful acquisition of knowledge by testimony. Testimony thereby differs from perception and memory because these considerations have no corollary in either perception or memory.

It might be argued that considerations of artfulness and dependability are comparable to considerations of illusion in perception and memory. It is equally true, for instance, that, if a subject is to acquire (noninferential) perceptual knowledge, then he must not be hallucinating. This parallel only holds, however, for the case of a speaker's dependability. The parallel does not hold in the case of artfulness, because the expression of something thought to be false is quite *intentional*. It is the relevance of another's intentions that renders testimony *fundamentally* distinct from perception and memory.

A speaker cannot *believe* something that he thinks is false. A speaker can *express* something that he believes is false. Insofar as our purposes in communicating are varied, insofar as a speaker can communicate with no intention either to inform or to deceive his audience, there seems to be no rational imperative constraining speakers to be artless. Even from the transcendental perspective, artful communication is rational. When we fail to acquire knowledge from perception, there is a sense that something has gone wrong. When we fail to acquire knowledge from communication, there is no sense that we have failed to communicate.[11]

---

[10] Dependability is not equivalent to reliability: it concerns a *particular* presentation.

The recognition of this distinguishing feature of testimony is reflected in our tendency to be skeptical of testimony. Our psychological attitude toward what we seem to perceive or remember is one of acceptance. Our psychological attitude toward what another tells us is to accept what we are told *only* given the presence of further background belief. These beliefs may relate to the testimony as a type. We distinguish types of content and types of speaker—an individual speaker can be thought of as a particularly narrowly defined type. These beliefs allow us to support acceptance with the judgment that the testimony is *credible*. On many occasions, our acceptance is also supported by a judgment that the particular content expressed is true. Our psychological disposition is to accept testimony only given suitable supporting background beliefs about the testimony's credibility and truth.

Given that a speaker's intentions in communicating need not be informative and given the relevance of these intentions to the acquisition of testimonial knowledge, I take this psychological disposition to be rational. It is doxastically irresponsible to accept testimony without some background belief in the testimony's credibility or truth. In the case of perception and memory, rational acceptance requires only the absence of defeating background beliefs. In the case of testimony, rational acceptance requires the presence of supporting background beliefs.[12]

---

[11] McDowell claims that there is a sense in which we have failed to communicate. There is a sense because the essential function of communication is the transmission of information. He justifies this claim by a speculative appeal to our evolutionary history. A bird's squawk can warn other birds of the presence of a predator and the assertoric use of language, McDowell suggests, is a descendent of such proto-communication. See his "Meaning, Communication and Knowledge" in Zak VanStraaten, ed., *Philosophical Subjects: Essays Presented to P.F. Strawson* (New York: Oxford, 1980), pp. 117-39. I should like to note three things. (1) Birds squawk more for display than to inform. (2) Group formation is plausibly the origin of human society and language. See Emile Durkheim, *The Elementary Forms of Religious Life: A Study in Religious Sociology* (London: Allen and Unwin, 1912); and Claude Levi-Strauss, *The Elementary Structures of Kinship* (Boston: Beacon, 1949). This suggests that the essential purpose of communication is the generation of the trust needed for any relationship. Jokes engender trust as much as informative truth telling. (3) It is probably better to side with Donald Davidson on this: "It would be good if we could say how language came into existence in the first place. These matters are, however, beyond the bounds of reasonable philosophic speculation"—"Three Varieties of Knowledge" in A. P. Griffiths, ed., *A.J. Ayer Memorial Essays* (New York: Cambridge, 1991), pp. 153-66, p. 157.

[12] One may put the *background* role of these beliefs this way. If an audience were to lack a dispositional belief in either the credibility of a given testimony or the truth

This demand of responsibility may be expressed as a criterion of justification. *An audience is justified in forming a testimonial belief if and only if he is justified in accepting the speaker's testimony.* Call this the *principle of assent.* The "justified" should be understood in Burge's narrow argumentative sense: a subject is justified if and only if there is an argument "available in the cognitive repertoire" of the subject to the truth of what is believed.

III. ASSENT AND ACCEPTANCE

I now want to argue that the principle of assent is compatible with the acceptance principle and consistent with its justification. I take the acceptance principle first.

The following distinction may be drawn. A justification connects an episode of believing with the truth of the proposition believed: it should state for what reasons a belief is held and it should determine that the proposition believed is, at least, likely to be true. These two aspects are conceptually separable: one may discriminate between the *justification of a belief* and the *justification of a proposition.* Thus, the distinction I should like to draw is between the justification supporting the proposition expressed by the speaker's testimony and the justification supporting the audience's acceptance of this proposition.

The acceptance principle draws an epistemic consequence from the following presumption. One may presume that (seemingly) intelligible testimony is true because one may presume that it is justified. (Read 'testimony' as 'presentation-as-true'.) This distinction between the justification of a belief and the justification of a proposition allows the following interpretation of this presumption. If a given testimony is justified, this justification epistemically supports the *proposition* that the testimony expresses. If an audience understands and accepts this testimony, this justification continues to support this proposition, now the object of the audience's belief. In this sense, justification is *preserved.* But *this* justification does not *justify* the audience's acceptance of the testimony.

To interpret the acceptance principle in this manner is to pursue an analogy suggested by Burge. "The Acceptance Principle," Burge states, "is clearly similar to what is widely called a 'Principle of Charity' for translating or interpreting others."[13] The "principle of charity" is

---

of the proposition it presented, the audience would not accept the testimony. Supporting points are made by Jonathan Adler, "Testimony, Trust, and Knowing," this JOURNAL, XCI, 5 (May 1994): 264-75.

[13] "Content Preservation," p. 487.

primarily a methodological rule of interpretation, which could be glossed: *one should optimize the agreement between oneself and another.* Given Donald Davidson's[14] view of interpretation, however, this methodological rule has the epistemic consequence that belief is by its nature veridical. This epistemic consequence could be formulated as the principle: *one is entitled, other things being equal, to presume that any given belief of a rational speaker is true.* Confronted with a sentence held-true, this principle does *not* entitle the interpreter *to accept* that *this* sentence *is true.* In forming beliefs as to *which* sentences held-true are true, the interpreter is guided by "hunches about the effects of social conditioning, and of course common-sense, or scientific, knowledge of explicable error."[15] How these beliefs are *individually* justified "is no easier to specify than to say what constitutes a good reason for holding a particular belief."[16]

On this re-interpretation, the acceptance principle does not state the conditions that justify belief. As such, it does not articulate an epistemic right to credulity. Thus interpreted, the acceptance principle might be better phrased: a person is entitled, other things being equal, to presume that those propositions presented as true and intelligible to him are justified. Or more simply, *a person is entitled to presume, other things being equal, that propositions expressed by intelligible testimony are justified.* To this principle the following claim about preservation should be appended: whatever justification in fact supports a testimonially presented proposition continues to support this proposition once accepted. This interpretation of the acceptance principle is consistent with the distinction between testimony and belief. Given the distinctiveness of testimony, one may presume that a *particular* proposition is justified only if one has reason to believe that the presenting testimony is credible (or one has independent reason to think that the proposition is true).

---

[14] "A Coherence Theory of Truth and Knowledge," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson,* pp. 307-19.

[15] Davidson, "On the Very Idea of a Conceptual Scheme," in his *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 183-98, here p. 196.

[16] Davidson, *Inquiries into Truth and Interpretation,* p. xvii. This quote is out of context but its guiding idea is not. The whole quote starts: "Understanding can be secured only by interpreting in a way that makes for the right sort of agreement. The "right sort" however is no easier...." Davidson's thought here is that the principle of charity cannot be made more specific than the rule glossed because whether the interpreter attributes error, whether or not he judges that in this case the sentence held true is not-true, is determined by his hunches, and so on. That is, the justification supporting these judgments is a matter of whatever "constitutes a good reason" for belief. It is not the entitlement stated.

Let me now turn to the justification of the acceptance principle. What I termed the principle of assent is consistent with Burge's justification of the acceptance principle, because the latter requires a strong assumption of rationality. This assumption, recall, is needed to support the crux premise in the justification of the acceptance principle. This is the premise that we are entitled to presume that another rational source of presentations-as-true is a source of true presentations. The *mere rationality* of another source entitles us to presume that the presentations-as-true of this source will be justified. It would entitle us to this presumption only if *being rational* could be identified with *being doxastically responsible.*

Our psychological disposition is to accept testimony only given the appropriate supporting beliefs. Given the unique character of testimony, this disposition is doxastically responsible. The principle of assent asserts that these reasons are needed for justified belief. This assertion is consistent with the justification of the acceptance principle because an audience must also be a rational source. One could possess the ability to understand presentations of propositional content only if one possessed the ability to present propositional content intelligibly. Thus, *if* there is a presumption that an interlocutor's presentations of content are backed by reasons, there should *equally* be a presumption that an interlocutor's reasons for accepting what is presented rationally back the belief formed through acceptance. That is, the entitlement articulated by the acceptance principle presupposes that the reasons accompanying acts of acceptance in fact justify our testimonial beliefs. This is captured by the principle of assent.

IV. THE HYBRID THEORY

I have re-interpreted Burge's acceptance principle and coupled it with a principle I termed the principle of assent. Before stating the resulting epistemological theory, let me get clear about terminology.

Let me introduce the term *'warrant'* in order to reserve the term *'justification'* for argumentative justifications, that is, in order to understand 'justification' in Burge's narrow sense. This is *not* to say that warrant is constituted by a reliable process. I would argue that being produced by a reliable process, or tracking the truth, is necessary to a belief's warrant, but I do not want to build this into the definition of 'warrant'. I want to understand 'warrant' generically as referring to any species of positive epistemic evaluation.[17] If, for instance, entitlements are a form of positive epistemic evaluation, then an entitled belief is a warranted belief. A positive epistemic evaluation which is broader than, and yet which encompasses, justification is needed because testimony allows that many of the things we believe may be supported by reasons that we have no access to and processes that are beyond us.

---

[17] As such, I am not understanding 'warrant' in Plantinga's sense of 'whatever is needed to make a true belief knowledge'. I allow for warranted true belief that is not knowledge.

With this definition in hand, I propose the following theory of testimony.

(A) An audience is warranted in forming a testimonial belief if and only if he is justified in accepting the speaker's testimony.
(B) If the audience is warranted in forming a testimonial belief, then whatever warrant in fact supports a speaker's testimony continues to support the proposition the audience believes.

Let me call this the *hybrid theory of testimony*. It is hybrid because it is composed of two principles, one *internalist* and the other *social externalist*. I named (A) the principle of assent; it articulates an internalist claim. It is internalist because the formation of a warranted testimonial belief requires a justifying argument that articulates the audience's reasons for believing. This argument must be based on the audience's own experiences; in the fundamental case, the audience's reasons for acceptance will be his experience of what testimonies are credible. Let me name (B) the *principle of warrant*; it articulates a social externalist claim. It is externalist because in accepting testimony an audience frequently believes propositions whose warrant could not possibly be equated with whatever justification the audience has for acceptance. This externalism is social because the existence of this warrant depends on there being other agents for whom this proposition is warranted. That is, the principle of warrant states the sense in which testimony preserves warrant.

Let me define the *extended warrant* of a proposition presented by testimony as the conjunction of the warrant possessed by the original speaker for believing this proposition together with any further warrant provided by the chain(s) of communicators. The presumption that intelligible testimony is warranted plays a role akin to that played by the principle of charity: an audience may presume that his *justification for acceptance* hooks onto the extended warrant of the proposition accepted. In the same way, an interpreter may presume that his reasons for taking a sentence to be true pick up the native's reasons. Our justifications for acceptance, as it were, refer to the further, more complete warrants that we take to support our testimonial beliefs.

The hybrid theory is hybrid in another sense: it can be seen as a combination of *reductionism* and *antireductionism*. Contrasting reductive and antireductive positions is one way of setting up the debate on testimony.[18] The reductionist would assert that it is our reasons for accepting testimony which provide the justification for our testimonial beliefs. The warrant of a testimonial belief is thereby *equated* with the warrant that is possessed for the associated judgment of credibility. This position is reductionist in that "testimonial warrant" thereby reduces to the warrant that we possess for beliefs deriving from other sources of knowledge. By contrast, the antireductionist would assert that a belief formed through credulous acceptance could be warranted *merely* because it is a testimonial belief. The judgment that a testimony is credible or that the proposition it expresses is true is thereby *fundamentally irrelevant* to the warrant of the associated testimonial belief.

Antireductionism is by far the most common epistemological position—on this definition, both Burge and McDowell are antireductionists—and reductionism has been severely criticized, most notably by C. A. J. Coady.[19] Both theories are grounded, however, in well-motivated epistemological ideas. Each may be seen as the product of an attempt directly to extend an epistemological theory that is successful, to whatever degree, for perception and memory. The reductive and antireductive positions then emerge as reactions of their guiding epistemological idea to the epistemic distinctiveness of testimony. That is, both positions may be represented as the product of those forces which have shaped the hybrid theory.

One could suppose that perception and memory are sources of knowledge because our perceptual and recollective faculties or mechanisms are reliable. This supposition could be extended. One could equally suppose that testimony is a source of knowledge because we possess a reliable *testimonial faculty*. This faculty would take as input the fact of a speaker uttering 'U' and expressing that *p*, the circumstances of the speaker's utterance, and any other relevant beliefs of the audience. As output, this faculty would produce in the audience the disposition to believe credible testimony and the disposition to disbelief otherwise.[20] Alternatively, one could suppose that perception and memory are sources of knowledge because an apparent perception or recollection provides defeasible reasons for belief. This supposition could again be extended. One could equally suppose that the mere fact of a speaker uttering 'U' and intelligibly expressing that *p* provides a prima facie reason for believing that *p*.

---

[18] For example, see Coady, "Testimony and Observation," *American Philosophical Quarterly*, x, 2 (1973): 149-55; Leslie Stevenson, "Why Believe What People Say?" *Synthese*, xciv (1993): 429-51, Elizabeth Fricker, "Telling and Trusting: Reductionism and Anti-Reductionism in the Epistemology of Testimony," *Mind*, civ, 414 (1995): 393-411. These positions are sometimes called Humean and Reidian, respectively.

[19] Anti-reductionism could, for instance, be attributed to Coady, Dummett, John Hardwig, P. F. Strawson, and M. O. Webb and Michael Welbourne.

[20] Fricker's posits the operation the following subjunctive sensitivity, which she calls *monitoring*: "It is true throughout of the hearer that if there were signs of

Both of these suppositions have a certain antecedent plausibility. The problem that testimony poses for both is that our psychological attitude is to be much more skeptical of what others tell us than what we seem to see or remember. I spent section II developing and defending this attitude. Recognition of this attitude, I suggest, provokes the following changes to the suppositions just made.

With respect to the supposition that we possess a reliable testimonial faculty, it may be the case that our dispositions to acceptance are in fact reliable, but whether or not this is the case is a matter of some concern for us. We know that speakers can be undependable and artful. We desire to avoid dumb credulity. Thus we possess quite extensive background beliefs to support any act of acceptance. Insofar as our acceptance of testimony is guided by this reasoning, it seems that any supposition of reliability must be based on the supposition that our reasoning is adequate. If the *process* of acceptance is essentially the *rational* process of assessing credibility, however, then it seems better to characterize the posited testimonial faculty in internalist terms. Reductionism is the position that results.

With respect to the supposition that mere intelligibility provides a reason for belief, it may be that there is some basic connection between intelligibility and truth, but it is matter of concern for us whether or not this connection holds between any given utterance and its truth. Again, we know that speakers can be undependable and artful. It seems plainly gullible just to accept something *merely* because it is intelligible. This is not to say that no such epistemological principle is tenable. But if such a principle is to remain tenable, it cannot claim that mere intelligibility provides a *reason for belief.* An alternative type of epistemic appraisal is needed. Burge introduces the notion of entitlement. Entitlement is *not* a reason for belief; the acceptance principle is not to be used as a premise in an argument to justify acceptance. The acceptance principle, rather, states that some-

---

untrustworthiness, she would register them and respond to them"—"Against Gullibility," in *Knowing from Words*, pp. 125-61, here p. 150. If she posited a correlative sensitivity to trustworthiness, then the ability to monitor could be understood in faculty terms.

times an audience is entitled to accept testimony credulously, without reason. This position is antireductionist.

The hybrid theory combines reductionist and antireductionist claims. The warrant of a testimonial belief cannot be equated with the warrant that is possessed for the associated judgment of credibility. The propositions that we come to believe through testimony may be supported by an extended warrant that is much greater than, and of quite a different nature to, the warrant supporting a judgment of credibility. But a judgment of credibility, or the judgment that the proposition a testimony presents is true, is not irrelevant to the warrant of the associated testimonial belief. No belief formed through credulous acceptance could be warranted; one needs reasons for accepting testimony.

V. COMMUNITIES OF KNOWLEDGE

Our justifications for acceptance can be knowledge supporting, but they need not be. Our acquisition of knowledge can depend on whether or not another possesses knowledge. The knowledge possessed by an individual can depend on the external social fact of whether or not warrant is possessed by others. How should this social nature of warrant be modeled?

To focus this question, let me consider a simple case of epistemic dependence. An original speaker S_1 sees that p and tells S_2 that p, S_2 tells that p to another, and so on until we get to S_n who tells that p to the audience A. Suppose that S_1 knew that p, that A was warranted in forming the testimonial belief that p, and that A did not have a knowledge supporting justification for acceptance. In this simple case, how should one characterize A's relationship to speakers S_1 to S_n? Within the context of the hybrid theory, the issue is how to interpret the principle of warrant, given a larger context of communicators. This question is independent of the hybrid theory, however: it should be addressed by any theory of testimony allowing that audiences can be epistemically dependent.

I want to consider two classes of answer to the question of A's relation to S_1 to S_n. On the first *transmission model*, S_1 to S_n form a *chain of testimony* and A knows that p if and only if there is knowledge in the chain. There are two options here, depending on whether A's knowing that p requires S_n has knowledge or merely that one of S_1 to S_n has knowledge. Burge's position fits the transmission model. He states: "If the recipient depends on interlocution for knowledge, the recipient's knowledge depends on the source's having knowledge as well."[21] This suggests the stronger claim that A's knowing that p

---

[21] "Content Preservation," p. 486.

requires S_n has knowledge. In a footnote, however, Burge immediately removes this implication: "Some chains with more than two links seem to violate this condition. But there must be knowledge in the chain if the recipient is to have knowledge based on interlocution" (*ibid.*, footnote 24). This makes the weaker claim that A's knowing that p requires one of S_1 to S_n to have knowledge. This weaker claim seems the right one. Suppose a teacher propounds a theory he does not believe. The teacher's rejection of the theory seems to imply that he does not have knowledge of it. If the theory is known, however, then it seems that the students could be in a position to acquire this knowledge.

The alternative is the *community model.* On this model, S_1 to S_n form a *community of knowledge* and A knows that p if and only if the community possesses knowledge.[22] Hardwig suggests that only the community model can explain scientific collaboration. He refers to an experiment,[23] recording charm events and measuring the life span of charm particles, that was conducted by *ninety-nine* physicists. This immense collaboration was required not simply because the experiment needed approximately 280 human/years to execute but also because a division of skills and knowledge was needed in order to execute it. Suppose that S_1 to S_n were the ninety-nine physicists and that p the experimental result. Hardwig[24] states:

> Unless we maintain that most of our scientific research and scholarship could *never*, because of the cooperative methodology of the enterprise, result in knowledge, I submit that we must say that p is known in cases like this. But if [A] knows that p, we must also say that someone can know "vicariously"—i.e., without possessing the evidence for the truth of what he knows, perhaps without even fully understanding what he knows. And this conclusion would require dramatic changes in our analysis of what knowledge must be. If this conclusion is unpalatable, another is possible. Perhaps that p is known, not by any one person, but by the *community* composed of [S_1 to S_n] (*ibid.*, pp. 348-49).

---

[22] The term 'community of knowledge' comes from M. Welbourne, "The Transmission of Knowledge," *The Philosophical Quarterly*, XXIX, 114 (1979): 1-9. He defines a community of knowledge in terms of *sharing* knowledge. If {A, B,...} share knowledge of the facts that {p, q,...}, then {A, B,...} *are* a community of knowledge. This community of knowledge is constituted by acts of telling that {p, q,...} as fact. I find this definition too reified. Academic philosophers may agree over few facts but surely constitute a community of knowledge. I use 'community of knowledge' to mean a social group constituted by acts of communicating about a certain topic.

[23] The article referred to is "Charm Photoproduction Cross Section at 20 GeV," *Physical Review Letters*, LI, 5 (1983): 156-59.

[24] "Epistemic Dependence," this JOURNAL, LXXXII, 7 (July 1985): 335-49.

Obviously, we want to allow that our scientific research can result in knowledge. Thus Hardwig claims that testimony presents the epistemologist with the following dilemma. An internalist requirement on knowledge is that a subject *S* can know that *p* only if *S* possesses evidence for the truth of *p*. The epistemic dependence experienced on certain occasions where testimonial knowledge is acquired then forces us either to abandon this internalist requirement—and thereby introduce "dramatic changes in our analysis of what knowledge must be"—or to regard this requirement as satisfied, in some sense, by the community of knowledge.

The hybrid theory avoids Hardwig's dilemma. The judgment that a testimony is credible provides a reason for thinking that the proposition it presents is true. Possessing a belief in the credibility of a testimony, the audience possesses evidence for the truth of the propositions it presents.[25] Burge would simply deny there is a problem: the acceptance principle and its justification do not introduce "dramatic changes."

My interest lies in the second horn of Hardwig's dilemma and the idea that a community can possess knowledge. Our linguistic practices certainly provide some support for this idea. We treat many scientific discoveries as *common knowledge*. For instance, that the Earth goes round the sun or that water is H_2O. If that *p* is common knowledge, one could state "we know that *p*" or speak of "our knowledge that *p*." In both these statements, there is an equivocation. The pronouns 'our' and 'we' could refer *distributively* to epistemic subjects qua individuals, or they could refer *collectively* to epistemic subjects qua members of some community. In this latter sense, the sentences 'we know that *p*' and 'our knowledge that *p*' refer to knowledge that is possessed by the community.

This raises the question of how a community could possess knowledge. Hardwig seems to suggest that it is *the community that knows*. Obviously, he would allow that *A* can acquire knowledge from S_n's testimony. Thus *A* could acquire knowledge even if none of S_1 to S_n has knowledge (and by the assumptions made *A* does not have a knowledge supporting justification for acceptance). I want to reject this notion of common knowledge. If *A* acquires knowledge, then at least one of S_1 to S_n must know that *p*. (Maybe S_n, the article writer who collates all the results.) This accords with the first horn of

---

[25] Admittedly, it is not the best possible evidence. But it is evidence in the ordinary way in which empirical generalizations are evidence for empirical predictions. And there could be no reasonable motivation for requiring that evidence be the best possible evidence.

Hardwig's dilemma. An internalist will require that the knower have some cognitive access to the warrant supporting what is known. A community could have access to the warrant supporting what is commonly known only if some community member had access to this warrant. That is, no one in a community can know unless someone does.

On the *transmission model*, S_1 to S_n form a *chain of testimony* and A acquires knowledge that p if and only if one of S_1 to S_n have knowledge. This condition should be understood recursively in the sense that S_n has knowledge if and only if one of S_1 to S_{n-1} has knowledge, S_{n-1} has knowledge if and only if one of S_1 to S_{n-2} has knowledge, and so on. Thus, it is *not* required that one of S_1 to S_n has access to, or awareness of, what warrant determines that p is known. On the *community model*, by contrast, S_1 to S_n form a *community of knowledge* and A acquires knowledge that p if and only if one of S_1 to S_n has access to, or awareness of, what warrant determines that p is known. (For example, and respectively, S_1 seeing that p and S_n collating all the results of the other scientists.[26])

There seems little to choose between these models; both can certainly explain Hardwig's case of collaboration. I think, however, that the community model is to be preferred because it provides a way of understanding the internalist "access condition" at the social level. To illustrate this I want to consider a case where I think that it is intuitive that knowledge is lost.

A warrant should determine at least that a proposition is likely to be true. We can directly see many ordinary empirical propositions—for instance, that it is raining outside—we can directly see to be true. The case of scientific theorems is quite different. We cannot directly judge whether or not a statement about charm particles is true. If we are to believe some such statement, a justification is required. In the illustration considered, the experimental result constituted the basis of such a justification. The case of mathematical theorems is similar. Whilst some mathematical or logical truths may be intuitively grasped, or "seen" to be true, most require a justification. Fermat's Last Theorem falls with the many: it is warranted only insofar as it is justified, that is, proved.

Suppose P. de Fermat did prove his Last Theorem. Let Fermat be the original speaker S_1 who tells S_2 his Last Theorem, S_2 then tells another, and so on until we get to S_n who tells Fermat's Last Theorem to A. And suppose that every acceptance of testimony was justified. Given these suppositions, it seems that A acquires knowledge.[27] Fermat can articulate the proof of his Last Theorem, therefore he has knowledge. At least one of S_1 to S_n has knowledge therefore on both the transmission and community models A acquires knowledge.

---

[26] Let me be clear. The formulation of these models in terms of necessary and sufficient conditions is compatible with the hybrid theory *only* given the background assumptions that (1) the audience's acceptance is warranted and (2) the audience does not possess a knowledge-supporting justification for acceptance.

Now consider the situation after Fermat's death. Suppose Fermat died after S_{n-1} told S_n Fermat's Last Theorem but before S_n told A. What is A's epistemic position given these suppositions? On the transmission model, it is no different. Once Fermat told S_2, S_2 has knowledge and, therefore, someone in the chain has knowledge.[28] On the community model, A's acceptance of S_n's testimony could still be justified by the belief that Fermat was a credible mathematician (and S_n an honest speaker) and thus A's testimonial belief could still be justified; but it could no longer count as knowledge, because the community that is S_2 to S_n no longer contains any members who could articulate a proof of Fermat's Last Theorem. The community, therefore, no longer contains any members who have knowledge. This bit of knowledge has been lost.

This is not to imply that lost reasons no longer warrant. It is plausible to suppose that, even if Fermat were to forget the proof of his Last Theorem, the fact that he did once articulate this proof is sufficient for his belief in his Last Theorem to count as knowledge. It is to imply that the conditions under which knowledge is lost to individuals differ from those under which it is lost to communities. If Fermat were to forget the proof of his Last Theorem (suppose he was suffering from senile dementia or some other ailment), then the community created by his testimony would lose access to this knowledge-supporting warrant. As such, the community could no longer be said to know Fermat's Last Theorem and this is just short-hand for saying that all those who know this proposition only in virtue of being a member of the community could no longer be said to know Fermat's Last Theorem.

---

[27] How difficult the proof was could be relevant. If the supposed proof was as difficult as the contemporary proof provided by Michael Wiles, then the audience would have a potential defeater for his judgment of credibility. That is, were he to know of the proof's difficulty he would have reason to think S_n's testimony non-credible. See Burge, "Computer Proof," note 9. Thus, for simplicity, let me suppose that the proof was as simple as Fermat's marginalia suggests.

[28] This is to consider only these members of the chain that are alive. If this restriction is abandoned, the situation is worse. It is worse because *once* somebody in the chain has knowledge, there will *always* be somebody in the chain who has knowledge.

It is hard to provide an account of how knowledge is lost on the transmission model. Knowledge clearly can be lost and I take it to be intuitive that it is lost in this instance. On the transmission model, it would not be lost. One may go further. It is plausible to suppose that there is a chain of testimony linking us to Fermat.[29] Thus, given the supposition that Fermat did prove his Last Theorem, its recent proof has made no change to this theorem's counting as knowledge. This seems wrong. I conclude that the community model provides a better conception of the social nature of testimonial warrant.[30]

VI. CONCLUSION

The case of Fermat illustrates a certain dynamic that exists between individual knowers and communities of knowledge. To conclude, I should like show how the hybrid theory provides a means of conceptualizing this dynamic.

According to the hybrid theory, the acquisition of knowledge requires the audience possess some *justification for acceptance.* But an audience's justification for acceptance will rarely recapitulate the *extended warrant* of the proposition accepted. A consideration of how the audience's justification for acceptance can be related to the extended warrant of the proposition accepted then provides a means of conceptualizing the epistemological engagement between an individual and the community.[31] This engagement has (approximately) the following pattern.

If the audience's justification for acceptance is *no more* than a judgment of credibility, then, ordinarily, it should provide no supplementary reason for the truth of the proposition accepted.[32] In this case, the audience acquires knowledge only insofar as the extended warrant possessed by the community is knowledge supporting. If the audience's justification for acceptance provides further reasons to believe that the proposition presented by the testimony is true, then these further reasons can supplement the extended warrant possessed by the community. Even if the audience's justification for acceptance were knowledge supporting, however, the prior extended warrant could still support the proposition accepted because the audience's justification would not, ordinarily, recapitulate this extended warrant.[33]

---

[29] In fact, there is a direct chain: Fermat's mathematical works are, what could be termed, *fossilized* testimony.

[30] I said that Burge holds the transmission model. This is clear in "Content Preservation." But in "Computer Proof" he has the following to say: "The conditions of dependence on previous members in the chain are complex. One's immediate interlocutor might perhaps not be knowledgeable. But the interlocutor must either pass on knowledge that reposes somewhere in the chain or, as a computer might, *otherwise indicate the existence of grounds for knowledge*" (note 9, my emphasis). Can one "otherwise indicate" historic warrants that are now inaccessible? If the answer to this is "no," then it seems that Burge's conception fits with the community model. One could interpret the italicized clause as a reference to a community possessed knowledge supporting warrant.

[31] This characterization of the relation of an individual to the community is not intended to be exhaustive. In particular, I am concerned with the social character of warrant, rather than content. The claim that content has a social character is made by Burge, "Individualism and the Mental," in Peter A. French, Jr., Thomas E. Uehling, and Howard K. Wettstein, eds., *Midwest Studies in Philosophy, Volume IV* (Minneapolis: Minnesota UP, 1979), pp. 73-121. And the claim that the relationship between warrant and content has a social character is made by Coady, *Testimony: A Philosophical Study*, pp. 169-73.

Knowledge is retained all the while the extended warrant that supports the proposition known is accessible to the community. To say that an extended warrant is accessible is to say that some set of community members, individually or conjointly, have access to or could articulate this extended warrant. Where the kind of knowledge retained is that which requires warrant to be some kind of justifying argument, paradigmatically mathematical and scientific knowledge, then the extended warrant the community must have access to must be such a justifying argument. Where the knowledge retained is everyday empirical knowledge, then the extended warrant the community must have access to could take one of three forms. (1) It could be the warrant that was originally associated with this piece of empirical knowledge. (2) It could be a justification based on this original warrant. (3) It could be a justification of the belief that the testimony expressing this piece of empirical knowledge is credible.[34]

---

[32] I say 'ordinarily' because in some cases the *only* warrant supporting a proposition is the judgment that the testimony presenting it is credible. For example, the audience may be a historian making a judgment about a piece of fossilized testimony. See condition (3) below.

[33] A dialogue at the conference of physicists would be an example here. I say 'ordinarily' because, rather than recapitulate the extended warrant, the audience's justification may defeat the extended warrant. For example, and speaking crudely, Copernicus defeated our warrant for the Ptolemic system.

[34] Consider the following situation. "Perhaps you and I and many others together map the coast of Australia: then I know by nontestimonial means that *this* bit has *this* shape; you know similarly *that* bit has *that* shape, and so on"—Alvin Plantinga, *Warrant and Proper Function* (New York: Oxford, 1993), p. 87. Suppose 'you and I and many others' are members of Captain Cook's crew. And consider the community of knowledge constituted by our composition of this map. In this community of knowledge, the knowledge that '*this* bit has *this* shape' satisfies (1). The knowledge that 'Australia is *this* shape' satisfies (2). Now suppose the ship returns to England, the map is accepted by the Royal Geographical Society and all the crew dies. In the community of knowledge created by the now dead crews' testimony the knowledge that 'Australia is *this* shape' satisfies (3).

In the case of (3) one could speak of the community judging this testimony to be credible.

Let me end on a metaphysical note. The hybrid theory makes the possession of testimonial knowledge a fundamentally social affair. It is social in the strong sense that knowledge can be supported by a warrant that requires group articulation. Both a speaker's and an audience's reasons may be relevant. The social dimension is not merely the external. I end with the thought that these cases are metaphysically fundamental. This, I think, is Robert Brandom's[35] point.

> If one *individualizes* the space of reasons, forgetting that it is a *shared* space within which we adopt attitudes towards *each other*—and so does not think about standings in the space of reasons as socially articulated, as potentially including the social difference of perspective between attributing and undertaking commitments, that is between your standing and mine—then one will not be able to understand knowledge as a standing in the space of reasons (*ibid.*, p. 904).

PAUL FAULKNER

University College London

---

[35] "Knowledge and the Social Articulation of the Space of Reasons," *Philosophy and Phenomenological Research*, LV, 4 (1995): 895-908.

RELIABILISM LEVELED*

Max knows that the 6 o'clock train goes all the way to Montauk. He just asked at the information booth in the station, where the staffers are highly competent and have direct access to the master timetable. But suppose that, instead of asking at the booth, Max had consulted an old, out-of-date schedule, and that many of the routes and times listed there had changed. Even if the schedule happened to say that the 6 o'clock goes to Montauk, Max would not know that it does. Whether Max knows or not seems to turn on the fact that information provided at the booth is extremely reliable, while information provided by the out-of-date schedule is not. Accordingly, one might hold that knowledge just *is* a reliably true belief, or, alternatively, a belief that results from a process that reliably produces true beliefs.

To adopt such a view is to endorse reliabilism with respect to knowledge. Here, I distinguish two versions of this position, which I call *neighborhood reliabilism* and *counterfactual reliabilism.* I then raise the question of what reliabilists can and must say about reflective or higher-level knowledge (that is, knowledge about one's knowledge or beliefs). It will emerge that both versions of reliabilism encounter serious difficulties in this connection, and must be regarded as unsatisfactory. These results suggest that the traditional justification requirement for knowledge cannot be supplanted by a reliability requirement, as many now are inclined to suppose.

I. RELIABILITY AND RELIABILISM

Reliabilism about knowledge takes different forms. To some extent, one can distinguish "reliable belief" accounts of knowledge from "reliable process" accounts. Roughly, according to the former, whether you know that P depends upon whether your *belief* that P is prone to error. According to the latter, whether you know that P depends upon whether the *process* that produced your belief that P is prone to error. Reliabilism about knowledge differs from reliabilism

---

*I wish to thank Jonathan Adler, William Alston, Stewart Cohen, Peter Klein, Hilary Kornblith, and Chad Mohler for helpful comments and discussion. I am especially indebted to Richard Feldman. Parts of this paper have their origins in a contribution to a symposium at the American Philosophical Association's Central Division meeting in 1995; a more recent draft was read at the Franklin and Marshall Colloquium on Metaphysics and Epistemology in 1999, and I appreciate the helpful discussion on that occasion. The title and much of the impetus for this paper are due to the work of Alvin Goldman.

0022-362X/00/9711/602–23

about justification, which seeks to explicate justification (rather than knowledge) in terms of reliability. The view I am considering involves no commitment to the notion that justification can be so understood.[1]

A principal motivation for adopting reliabilism is the idea that truth is the ultimate epistemic norm, and that besides truth, there are no others. We want what we believe to be true. Moreover, because truth is so important, we want our beliefs to be not just adventitiously true, but *securely* true. Knowledge, then, is securely true belief, and knowledge is compromised to the extent that there is falsehood or the possibility of falsehood. The greatest security one could have would be certainty or infallibility. A belief would be certain if it could not possibly be wrong. That is, in any possible world in which you believe X, X is true:

(1) Necessarily, $B(X)$ implies X.

and, equivalently:

(2) Necessarily, $\neg X$ implies $\neg B(X)$.

Similarly, a belief-forming process would be infallible if it could not possibly go wrong, and this conception can be spelled out along the lines of (1) and (2).

But such logical invulnerability to error is often an unattainably high standard. Whenever someone believes a contingent proposition as the result of an inductive inference, it is logically possible for the belief she arrives at to be wrong. Hence, it seems that knowledge must demand something less than certainty or infallibility—that is, something less than truth in any logically possible situation. Knowledge may require reliability instead. What is crucial is that reliability, whether it pertains to beliefs or to processes, need not be total or absolute. Something may do Y reliably, even if it would fail to do Y under very extreme or extraordinary conditions. An alarm clock may be reliable, despite the fact that it would not ring if the power went out, or if the ceiling collapsed on it. Accordingly, reliability as it pertains to knowledge may be understood as truth in all situations that could arise, *except* for extraordinary or outlandish ones that we do not, and need not, care about. Thus, perception, or particular beliefs formed by perception, may count as reliable, despite the fact that perception, or particular beliefs arrived at by perception, can go wrong under certain circumstances.

---

[1] Rather, the reliabilist I have in mind denies that there is any independent, substantive justification condition for knowledge; see below. It may well be, though, that the arguments I shall present also tell against reliabilism about justified belief. I should mention that there is a hybrid of justificationist and reliabilist accounts, sometimes called the "reliable indicator theory." According to such a view, S knows that P only if S's belief that P is based on evidence E, where E is a reliable indicator of P. I shall not discuss this approach here.

This understanding of epistemic reliability can be sharpened somewhat, as follows. Think of possible worlds as being farther away from the actual world to the extent that they differ from the actual world. A belief is reliable just in case it turns out to be true whenever it is held in a neighborhood N of worlds not too far away from the actual world; a process is reliable just it case it yields (mostly) true beliefs in a neighborhood N of worlds not too far away from the actual world. Proceeding in this vein, the reliabilist may set out as a condition of knowledge:

(3a) In N, $B(X)$ implies X.[2]

or:

(3b) In N, $B(X)$ by process P implies X.

or:

(3c) In N, all or nearly all beliefs $B(X_i)$ formed by process P are such that $B(X_i)$ implies X_i.

Assume, at least for now, that the boundary of N is fixed. In particular, it does not vary with the content of the proposition believed, the process of belief formation, or pragmatic considerations.

(3a)-(3c) invoke one conception of reliability, but there are others. And, clearly, there will be different versions of reliabilism corresponding to the different conceptions of reliability one might adopt. Theorists such as David Armstrong, Fred Dretske, Robert Nozick, and Keith De Rose have held that knowledge requires satisfaction of a co-variation or tracking condition:

(4) If X were false, then $\neg B(X)$.[3]

---

[2] Note that we do not want the converse condition for knowledge, namely, that in N, X implies $B(X)$.

[3] Like (3), the tracking condition is subject to modification. (4) as stated corresponds to (3a) above. An analogue to (3b) would be:

(4b) If X were false, then $\neg(B(X)$ by process $P)$.

where P is the process you actually used to arrive at your belief that X. Proponents of the tracking condition, like Nozick and De Rose, have in fact been led to include some reference to methods or processes in their accounts. But that has proved to be something of a quagmire; it is unclear how the relevant method or process is to be specified, and it is difficult to say exactly how reference to the process should figure in the

A belief might meet a version of condition (3) as stated, yet not meet condition (4). Here is why. Consider (3a) and (4). Given the standard Lewis-Stalnaker semantics for counterfactuals, instances of (4) are true if and only if the consequent is true in the closest possible world (or worlds) in which the antecedent holds. Which world that is, and how far it is from the actual world, depends upon what the value of X is. To illustrate, consider two propositions I know:

(5) I am not carrying a pen with blue ink.
(6) The Earth is not governed by cows.

Condition (4) makes my knowledge of (5) depend upon what I would believe in the closest possible world in which I was carrying a pen with blue ink, and it makes my knowledge of (6) depend upon what I would believe in the closest possible world in which the Earth was governed by cows. The closest $\neg(5)$ world—one in which I happen to be carrying a pen with blue ink—need not be very different from the actual world, in which I happen to be carrying a pen with black ink. That is, the closest $\neg(5)$ world is very close to the actual world, and we may assume that it falls within the boundary of N. A possible world in which the Earth was governed by cows would, however, be very different from the actual world, so different that it might well lie beyond the boundary of N. We may suppose that my belief as to whether the Earth is governed by cows is always true inside N, and that I thereby satisfy condition (3a). It remains open that I am wrong about this matter in the nearest world where cows do govern, contrary to (4). Thus, satisfaction of condition (3a) with respect to a given proposition does not guarantee satisfaction of condition (4) with respect to that proposition.[4]

In light of the foregoing, we can distinguish two different versions of reliabilism. The first, which I shall call *neighborhood reliabilism* (NR), is the view that you know X, if and only if you have a true belief that X which satisfies some version of (3). The second, which I shall call *counterfactual reliabilism* (CR), is the view that you know X, if and only if you have a true belief that X which satisfies some version of both (3) and (4).

It is important to note what requirements these theories do *not* impose on knowledge. First, NR and CR eschew the traditional claim

---

account. For further comment on (4b), see footnote 18. I cannot think of any version of the tracking condition that corresponds neatly to (3c).

[4] Another way to make the point: suppose you know the proposition that the actual world is in N. Evaluation of whether that belief satisfies (4) requires considering the truth of what you believe in worlds outside N.

that one could not know a proposition without having justification for believing it. Let us say that to be justified in believing a proposition means something like having beliefs or nondoxastic states that provide good reasons for believing that proposition. Reliabilists who reject the justification condition sometimes argue as follows. The justification condition is independent of the reliability condition only if holding beliefs for good reasons is not necessarily truth conducive. In that case, the pursuit of rational belief and the pursuit of true belief could, at least in principle, diverge. But truth is the sole and ultimate epistemic value, so any substantive, independent justification condition has no place among the requirements for knowledge.[5]

That is one rationale for giving up the justification condition. A highly advertised benefit of doing so is that it immediately allows us to be rid of at least some forms of skepticism. According to a very familiar line of skeptical argument, I have no justification for believing that I am not a thoroughly deceived brain in a vat, and consequently I fail to know that I am not. Since I fail to know that I am not thoroughly deceived in this way, I have no knowledge of the external world. A reliabilist, or anyone else who rejects the justification condition, will balk at the first step. She may concede that I have no justification for believing that I am not a brain in a vat. Unless knowledge requires justification, however, it does not follow that I fail to know that I am not a brain in a vat, and the skeptical argument is blocked. I shall return to this issue later.

A second point is that reliability theorists typically reject various "higher-level" requirements on knowledge, such as:

(7) $K(X)$ entails $K(K(X))$.

and:

(8) $K(X)$ entails $K(R(X))$.[6]

The reliabilist has a number of motivations for denying (8), and, therefore, (7).

For one thing, a reliable belief or process is truth tropic regardless of whether it is *known* to be so. Hence, a requirement that, beyond being reliable, a belief or process has to be known to be reliable would not do anything to foster the epistemic norm of truth. Such a requirement, then, has no place in an account of knowledge.

---

[5] For an argument in this spirit, see Alvin Goldman, *Epistemology and Cognition* (Cambridge: Harvard, 1986), pp. 97-103.

[6] Read '$R(X)$' as 'S is reliable about X'. When I say 'S is reliable about X', that locution is meant to stand in for whatever more specific reliabilist condition one might adopt. I shall assume, as a matter of terminology, that you are reliable about X only if you believe X and X is true. So, under this regimentation, $R(X)$ entails $B(X)$ and X. Moreover, for the reliabilist, $R(X)$ becomes equivalent to $K(X)$, and (7) becomes equivalent to (8).

Next, from the reliabilist standpoint, you know *X* in virtue of being reliable about *X*. *X* is one fact about the world; whether you are reliable about *X* is another, different fact about the world. You are reliable about *X* if, in your environment, you are suitably related to *X*. You may be so related to *X* without being so related to *R*(*X*). Hence, you can be reliable about *X* but not reliable about *R*(*X*). You would, then, know *X*, but not know that you are reliable about *X*.

In addition, (8) as stated apparently requires an endless sequence of beliefs *B*(*R*(*X*)), *B*(*RR*(*X*)), *B*(*RRR*(*X*)), and so on, such that one is reliable about *R*(*X*), *RR*(*X*), *RRR*(*X*), and so on. That is unattractive, certainly. And, if each of these beliefs has to be generated by distinct or additionally complex mechanisms or procedures, the regress is utterly vicious.[7]

Finally, the denial of (8) must be independently plausible, if reliabilism is to withstand one of the main criticisms which has been directed against it. This criticism, due to Laurence Bonjour[8] and Keith Lehrer,[9] is presented by way of various examples. Here is Bonjour's well-known "Norman" case:

> Norman, under certain conditions which usually obtain, is a completely reliable clairvoyant with respect to certain kinds of subject matter. He possesses no evidence or reasons of any kind for or against the general possibility of such a cognitive power, or for or against the thesis that he possesses it. One day Norman comes to believe the President is in New York City.... In fact the belief is true and results from his clairvoyant power under circumstances under which it is completely reliable (*op. cit.*, p. 62).

Bonjour maintains that Norman does not know in this instance, and concludes that reliabilism is untenable. Even among those who find this example convincing, there is some difference of opinion as to what defect in reliabilism it brings to light. But one prominent reaction is that Norman does not know the president is in New York because he fails to satisfy some higher-level requirement on knowledge.[10] If Norman had verified that the deliverances of his clairvoyant power reliably report how things are, then he would know that the president is in New York. But, since he lacks this higher-level knowledge, Norman lacks the first-level knowledge that the president is in New York.

---

[7] See Goldman, "Naturalistic Epistemology and Reliabilism," in Peter A. French, Theodore E. Uehling, and Howard K. Wettstein, eds., *Midwest Studies in Philosophy*, Volume XIX (Minneapolis: Minnesota UP, 1994), pp. 310-12. He claims that (7) and analogous principles create a vicious regress for nonreliabilist theories.

[8] "Externalist Theories of Empirical Knowledge," in French, Uehling, and Wettstein, eds., *Midwest Studies in Philosophy*, Volume V (Minneapolis: Minnesota UP, 1980), pp. 55-73.

[9] *Theory of Knowledge* (Boulder: Westview, 1990).

[10] This is Bonjour's own conclusion (pp. 62-64; see also Lehrer, p. 165).

I am not sure exactly what to make of this discussion, since I tend to lose my bearings when clairvoyance is so much as mentioned.[11] I do think, however, that this much is relatively clear: it is possible for one to be reliable about *X* without knowing that one is. According to reliabilist accounts of knowledge, simply being reliable about *X* is supposed to be sufficient for knowing that *X*. If knowing that you are reliable about *X* is an additional necessary condition for your knowing *X*, then these reliabilist accounts are mistaken.[12]

To the reliabilists I have in mind, the big picture looks like this. We know a lot about the world. We are reliable about things we believe on the basis of perception, so we know all kinds of things by seeing, hearing, and so forth. In addition, we are competent reasoners, at least about relatively elementary matters. We are reliable when we arrive at beliefs by induction or deduction, and these beliefs also count as knowledge. If we think about it, we probably believe that we are reliable about the things we believe on the basis of perception and inference. It is not apparent, however, whether these higher-level judgments are themselves reliable. If and when they are not, then we fail to know that we know. But, says the reliabilist, there is nothing wrong with that.

---

[11] It seems to me, though, that a less far-fetched example to the same effect could be given. A few people have perfect pitch; they can immediately identify what absolute pitch they are hearing. I take it that whether someone does in fact have perfect pitch is not immediately apparent to that individual. To tell whether you have perfect pitch, you need to be tested. Let us suppose, then, that many people believe that they have perfect pitch, although they really do not. Now consider Norma, who has perfect pitch, but has never checked whether she does. She hears a tone, which she correctly takes to be an 'A'. Does she know that the tone is an 'A'?

[12] I should note that Goldman distinguishes "basic psychological *processes*" (my emphasis) of belief formation from *methods*, which are "various sorts of algorithms, heuristics, and learnable methodologies"—*Epistemology and Cognition*, p. 93. He also writes: "They are not part of the fixed, native architecture of the cognitive system. This is sufficient grounds for their not being basic processes" (p. 366). Goldman suggests that it is not enough for a method to be reliable to pass muster; it has to be acquired by a process or method that has higher-order reliability. That is, the first-order method needs to be acquired by a procedure that reliably leads to the acquisition of reliable methods (pp. 91-93). To this extent, a reliabilist account of knowledge modeled on Goldman's account of justification might recognize a higher-order requirement in the case of non-native, acquired belief-forming methods. But as Goldman himself acknowledges, the distinction between processes and methods is difficult to draw (p. 92 note 11). For example, the capacity to see requires maturation and stimulation in order to develop, but Goldman does not want perceptual beliefs to be subject to higher-level requirements. A further qualm is that it is hard to see how instituting such a requirement for methods can be squared with the thought that all we should ask for from beliefs or ways of forming beliefs is reliable truth.

II. COUNTERFACTUAL RELIABILISM IS TOO STRONG

I think the right picture is somewhat different. I agree that we know a lot about the world. I also agree that, when we form various beliefs, those beliefs or belief-forming processes may be reliable without our knowing that they are. Nevertheless, we often know that our beliefs about the world are true, not false. On some occasions we know, further, that we are reliable about what we believe. And, finally, there are many instances when we know that we know. It is incumbent upon reliabilists to get these facts about higher-level knowledge straight. I do not think they can.

Let me begin with CR. My treatment of it will be relatively brief, in part because I have discussed the status of (4) as a condition for knowledge elsewhere. (4) was the subject of intense scrutiny some years ago, and serious problems with it came to light. For one thing, it leads to unacceptable failures of what is known as the *closure principle for knowledge*:

(9) If $K(P)$ & $K(P \text{ entails } Q)$, then $K(Q)$.

Also, there are unexceptionable cases of inductive knowledge for which (4) does not hold, so (4) is too strong.[13] There is an additional difficulty: (4) is hard to square with a satisfactory account of higher-level knowledge.[14]

Consider the following example. You see your long-time friend Omar, who is a perfectly decent and straightforward sort of person. Noticing his shiny white footwear, you say, "Nice shoes, Omar, are they new?" Omar replies, "Yes, I bought them yesterday." I think the following things are true:

(10) You know Omar has a new pair of shoes.

---

[13] See my "Tracking, Closure, and Inductive Knowledge," in S. Luper-Foy, ed., *The Possibility of Knowledge: Nozick and His Critics* (Totowa, NJ: Rowman and Littlefield, 1987), pp. 197-215.

[14] What follows is a more emphatic version of a criticism I originally directed against Nozick's account of knowledge (*ibid.*, p. 203). In his "Solving the Skeptical Problem," *The Philosophical Review*, CIV (1995): 1-51, De Rose has defended something very like (4) (although, to be precise, De Rose's view is not that (4) is actually a condition for knowledge per se, but works instead to set the standards for correct knowledge attributions). De Rose takes note of the sort of problem I raised, for which he offers no solution (p. 22). He nevertheless retains confidence in a version of the counterfactual requirement, though on what basis I cannot see. For additional discussion of counterfactuals and knowledge, see my "Subjunctivitis," *Philosophical Studies* (forthcoming).

(11) You know that your belief that Omar has a new pair of shoes is true, or at least not false.[15]
(12) You know that your belief that Omar has a new pair of shoes is reliable.[16]
(13) Other things being equal, you know that you know that Omar has a new pair of shoes.

(13) and (12) depend upon (11), and (11) is certainly compelling in its own right. (13) entails (11), because to know X entails that you have a true belief that X. (12) entails (11), if your being reliable about X is understood to entail that your belief that X is actually true.

The prime question at this point whether CR can accommodate (11). Does CR allow you to know that your belief that Omar has a new pair of shoes is true, not false? In particular, does (11) meet the condition set by (4)? I think the answer is no. I shall try to say why somewhat loosely and informally, and then more carefully.

(4) requires that, if you know X, then you would not believe X, if X were false. As things actually are, you believe that your belief that Omar has new shoes is not false. What if it were? If somehow your belief that Omar has a new pair of shoes were false, you would still believe that your belief was true, not false. The alternative is hard to fathom. It is difficult to conceive of your not believing that something you believe is true, whenever the matter happens to cross your mind. So, if your belief that Omar has new shoes were false, you would still believe that your belief was true, not false. You thereby fail to satisfy condition (4). According to CR, then, you do not know that your belief that Omar has new shoes is true, not false.

Let me now go through the same point a little more carefully. Let O = 'Omar has new shoes'; $B(O)$ = 'You believe Omar has new shoes'; $B(\neg(B(O) \,\&\, \neg O)$ = 'You believe that you do not believe falsely that Omar has new shoes'.

(14) In the actual world, $B(\neg(B(O) \,\&\, \neg O)$.

Consider the nearest possible world W in which your higher-level belief is false, that is,

(15) In W, $\neg(\neg(B(O) \,\&\, \neg O))$:

---

[15] Note, in addition, that any proposition X itself entails that a belief that X is not false: X entails $\neg(B(X) \,\&\, \neg X)$. So, to say that you know that Omar has new shoes, but you fail to know that you do not believe falsely that Omar has new shoes would be to reject the closure principle for knowledge.

[16] Alternatively: you know that the process by which you arrived at that belief is reliable.

Simplifying,

(16) In W, $((B(O) \& \neg O)$.

From (16),

(17) In W, $B(O)$.

If you believe O, you believe that you do not falsely believe O (see above). Hence,

(18) In W, $B(\neg(B(O) \& \neg O)$.

Given (15) and (18),

(19) In W, $(B(O) \& \neg O) \& B(\neg(B(O) \& \neg O))$.

That is, if your belief that O were false, you nevertheless would believe that it was not false. So, according to (4), you fail to know that you do not believe falsely that O. It follows, I take it, that you also fail to know that your belief is true.[17]

Note that the argumentation here is quite general. (4) makes it impossible for you to know that *any* of your beliefs is true, not false.[18] Thus, it also seems to exclude your knowing that any of your beliefs is reliable, or that any of your beliefs is knowledge. That is going far too far, in my book.[19] So, let us set CR aside, and move on to NR.[20]

III. NEIGHBORHOOD RELIABILISM IS TOO WEAK

NR makes knowledge out to be true belief that satisfies some version of (3). The differences among the various formulations, and the forms of NR they give rise to, can be quite significant. The trend these days seems to be toward reliable-process theories, which run along the lines of (3c) rather than (3a) or (3b). Hence, in what follows, I shall consider the variant of NR that incorporates (3c), although what I shall have to say carries over to the other forms of NR as well.[21] As I have set things up, NR is generally weaker than CR, since NR forgoes condition (4) while CR includes it. I have maintained that CR is too strong. It stints us higher-level knowledge we actually have. Now, I shall argue that NR is too weak. It would allow us higher-level knowledge we do not have.

---

[17] Unfortunately, this maneuvering is clumsy and roundabout. I did it this way in part to avoid complications about counterfactuals with disjunctive antecedents. Such technical hazards will arise any time one tries to apply (4) to knowledge of a conjunction. So far as I am concerned, that itself is good reason to be suspicious about (4).

[18] Very much the same goes for the modification of (4) mentioned in footnote 3, namely, (4b). Suppose you believe O by some particular process P. It should be possible for you to know that you do not falsely believe O by process P. That knowledge requires $B(\neg(P(O) \& \neg O)$, where '$P(O)$' now stands for the proposition that you believe O by way of process P. Suppose that your belief were false, that is, $(P(O) \& \neg O)$. Make the further, plausible assumption that your path from believing O by process P to a belief in the falsity of the conjunction would remain the same. Then, your belief in the falsity of the conjunction fails to satisfy condition (4b). In other words, you cannot know that you do not falsely believe O by process P.

[19] This argument also cuts against reliabilism that accepts (4) without (3).

[20] A proponent of CR might attempt to surmount this difficulty by adding to her account a stipulation that you know any propositions you deduce from other propositions you know. This step is fraught with difficulties, however. The revised account may be too weak, and fail for the same reason that NR does (see below). Moving in this direction would also exacerbate other problems facing CR, such as those raised by Saul Kripke in widely heard but unpublished lectures.

Once again, I shall proceed by considering an example, namely, the "gas-gauge case," due in its original form to Michael Williams.[22] Williams describes himself driving a car with a working, highly reliable gas gauge. Williams does not know, however, that the gauge is reliable. Let us stipulate that he has never checked it,. he has never been told anything about its reliability, and he does not even have any background information as to whether gauges like his are likely to be working. He never takes any special steps to see whether the gauge is going up or down when it ought to be. Rather, without giving the matter a second thought, Williams simply goes by what the gauge says. The gauge reads 'F', and Williams believes that his gas tank is full. According to NR, he *knows* that his tank is full. He has this knowledge because his belief results from a reliable process, that is, going by a well-functioning gas gauge. But Williams does not *know that he knows* that his tank is full. To have this higher-level knowledge, he would need to know that the gauge reliably registers the level of gas in his tank, and we have stipulated that he has no such information.[23]

No one should be distracted by the details of the case. Perhaps it is hard to imagine how, in the situation described, Williams could fail to have *any* background information about the reliability of his gas gauge. Then again, one might doubt whether the supposedly reliable process Williams uses is appropriately identified as "going by a well-functioning gas gauge." But what is essential here is just the claim that a subject might arrive at a belief by some reliable process, appropriately identified as *P*, yet not know that *P* is reliable. If particular features of my example create difficulties for you, let reading the gas gauge be a proxy for any such process *P*.

---

[21] For a somewhat different taxonomy of reliabilist positions, and for some arguments that would favor what he calls a "global" process theory like (3c), see Goldman, *Epistemology and Cognition*, pp. 43-49. I am not sure that Goldman's official scheme has room for NR as I construe it, since he seems to assume that what he calls a "relevant alternatives" theory is a broadening of a "pure subjunctive theory" (p. 45). As I would put it, Goldman assumes the reliabilist accepts (4), and the open question is then whether to add some version of (3). I have set things up the other way around.

[22] *Unnatural Doubts* (New York: Oxford, 1991), p. 347.

[23] Williams himself says: "No item of knowledge guarantees full knowledge of its own reliability conditions: indeed this is just another way of stating the externalist point that knowing does not guarantee knowing that one knows" (*ibid.*, p. 348).

Now consider another driver, whom I shall call Roxanne. She is like Williams, in that she believes implicitly what her gas gauge says, without knowing that the gauge is reliable. But she undertakes the following, admittedly curious, procedure. She looks at the gauge often. Not only does she form a belief about how much gas is in the tank, but she also takes note of the state of the gauge itself. So, for example, when the gauge reads 'F', she believes that, on this occasion, the tank is full. She also believes that, on this occasion, the gauge reads 'F'. Moreover, Roxanne combines these beliefs; she believes:

(20) On this occasion, the gauge reads 'F' and *F*.

Certainly, the perceptual process by which Roxanne forms her belief that the gauge reads 'F' is a reliable one. By hypothesis, her belief that the tank is full is also reached by a reliable process. Hence, there seems to be no good reason to deny that Roxanne's belief in (20) is the result of a reliable process, and the reliabilist will say that she knows (20).[24]

Now, it is a completely straightforward logical consequence of (20) that:

(21) On this occasion, the gauge is reading accurately.

Assume that Roxanne deduces (21) from (22). Deduction is certainly a reliable process, so there is no loss of reliability at this step. Consequently, it seems that Roxanne must be credited with knowing (21).

---

[24] What I have said may provoke disquiet of the following sort. You might believe that *A* by way of a reliable process, and believe that *B* by way of a reliable process, yet your belief that (*A* & *B*) might not be the result of a reliable process. For one thing, the chances of error with respect to *A* may be just barely acceptable and the chances of error with respect to *B* may be just barely acceptable. The chances of error with respect to the conjunction may then come out to be higher than the chances of error with respect to the conjuncts, and so the belief in the conjunction may not count as reliable. Another problem is that the process by which one comes to believe that *A* may interfere with the process by which one comes to believe that *B*, or vice versa. For example, one might use one instrument to measure for *A*, and another to measure for *B*, yet the operation of the second instrument may disrupt the functioning of the first. So, I think it is true in general that $R(X)$ and $R(Y)$ do not entail $R(X \& Y)$. But it seems implausible that such a failure occurs in the situation under discussion.

She knows this, supposedly, despite the fact that she has no independent information at all about the reliability of the gauge—whether it is broken or likely to be broken, whether it is hooked up properly, and so on. She just looks at the gauge and immediately believes what it says.

Let us say further that Roxanne does this over and over again for a good while. At various times *t*, she looks at the gauge, which reads '*X*', and forms the belief that, on occasion *t*, the gauge says '*X*' and the tank is *X*. Given what has already been said, she comes to know that, on each of these occasions, the gauge was reading accurately. Then, putting these pieces of information together, Roxanne concludes by induction:

(22) The gauge reads accurately all the time.

Reliabilists generally accept that induction is a reliable belief-forming process. So they should concede that the transition from various beliefs like (21) to (22) is knowledge preserving. We can add that, from (22), Roxanne infers:

(23) The gauge is reliable.

If Roxanne knows that the gauge is reliable, then she can presumably deduce and know that the process by which she comes to believe that her gas tank is full is a reliable one. And with just a little more deduction, Roxanne can come to know that she knows that her gas tank is full.[25]

This extraordinary procedure, which I shall call *bootstrapping*, seems to allow one to promote many, if not all, of one's beliefs that were formed by reliable processes into *knowledge* that those beliefs were formed by reliable processes.[26] I assume that bootstrapping is illegitimate. Roxanne cannot establish that her gas gauge is reliable by the peculiar reasoning I have just described. The challenge to NR is that it may go wrong here. On the face of things, it does improperly ratify bootstrapping as a way of gaining knowledge.[27]

---

[25] There is a noteworthy similarity between the problem I am raising for reliabilism and a difficulty that faces co-variational accounts of semantic content. The former is an attempt to analyze knowledge in terms of the co-occurrence of beliefs and the facts; the latter is an attempt to analyze semantic content in terms of the co-occurrence of linguistic or mental tokens and the facts. A difficulty for the latter is that, if tokens of *R* co-occur with distal events *X*, they also co-occur with the proximate process *P* that links *X*'s with *R*'s. So, the result would be that the *R*'s just as well represent *P* as *X*. See Fred Dretske, "Misrepresentation," reprinted in William Lycan, ed., *Mind and Cognition: A Reader* (New York: Oxford, 1992), pp. 129-43. The parallel problem I am raising is that, if the output of some belief-forming process *P* reliably co-occurs with *X*, that output also reliably co-occurs with instances of a reliable belief-forming process, namely, *P* itself. Thus, the result seems to be that *P* itself can give one knowledge that one is linked to *X* by way of a reliable process, that is, *P* can directly give one knowledge that one knows that *X*. The general lesson to be drawn is that co-occurrence can obtain for a variety of reasons, and is of questionable value in understanding relations like representation and knowledge.

To sort this out, we need to get clear about what the defect in bootstrapping is. I shall consider some possible diagnoses, in a point-counterpoint format:

Point. It is obvious what Roxanne did wrong. Suppose the gauge had not been reliable. Roxanne would have gone through exactly the same procedure, and reached the false conclusion that her gauge was reliable. Hence, she does not know that the gauge is reliable.

Counterpoint. Say that if you like. But then you are really assuming that knowledge has to satisfy condition (4). In other words, you are abandoning NR in favor of CR, and we have already seen that CR fails.

Point. The problem with Roxanne's procedure is that it could not have possibly yielded any other result other than the one it did, namely, that the gauge is reliable.

Counterpoint. It is not clear that the putative defect really is one. The process by which I know I am conscious when I am is surely a reliable one, yet that process could not return a verdict other than that I am conscious.[28]

Point. Roxanne does not know that the gauge is reliable, because she reaches that conclusion by bootstrapping, and bootstrapping itself is an unreliable process. After all, you can apply bootstrapping to a great many underlying processes, some reliable, some not. Every time, though, bootstrapping will tell you that the underlying process is reliable. When the underlying process is unreliable, bootstrapping will yield the false belief that the underlying process is reliable. So, bootstrapping itself often generates false beliefs, and must be considered unreliable.

---

[26] For definiteness, let us say that bootstrapping is the procedure that leads to beliefs, like (23), about the reliability of its underlying process. Bootstrapping may require that one be able to identify appropriately the underlying process by which one has arrived at a particular belief, and sometimes one may not meet that condition.

[27] Richard Fumerton independently arrived at a somewhat similar point—*Metaepistemology and Skepticism* (Lanham, MD: Rowman and Littlefield, 1995), pp. 178-79.

[28] Something like this objection is raised by William Alston, *The Reliability of Sense Perception* (Ithaca: Cornell, 1993), p. 17. He does want to allow, however, that introspection counts as a reliable process (p. 139).

Counterpoint. So far as I can tell, this is the best answer for the reliabilist to give, but it still seems problematic in various respects. My first worry is a methodological one. The response under consideration identifies the process leading up to Roxanne's belief in (23) as a token of a relatively wide process type, roughly *bootstrapping in general.* The failures of other tokens of this broad type are supposed to discredit Roxanne's undertaking, and the belief that results from it. Such a response is dissatisfying because it is generally possible to find a process type under which a given process token falls, such that the token counts as reliable, just as it is generally possible to find another process type under which the token falls, such that the token counts as unreliable. There is no agreed-upon, principled way of identifying which is the proper process type to consider in evaluating the epistemic status of a particular belief.[29]

The reliabilist herself may encounter some embarrassment on just this score. One might describe Roxanne's initially going by her gas gauge as "forming a belief by an *unauthenticated* process," that is, a process one does not know to be reliable. By parallel with the argument just given against bootstrapping, one might say that, if one forms beliefs by an unauthenticated process, often enough the token process one employs will be unreliable, and one's belief will be false. Hence, forming a belief by an unauthenticated process is generally unreliable, and beliefs that result from unauthenticated processes do not count as knowledge. Of course, the reliabilist rejects the demand that a process must be known to be reliable if it is to be a source of knowledge. She then has to find some basis for allowing the criticism of Roxanne's bootstrapping to stand, while rejecting the corresponding criticism of Roxanne's forming a belief by an unauthenticated process in the first place. Otherwise, the reliabilist's charge of guilt by association may boomerang.

Suppose we set this scruple aside, and accept that Roxanne's way of reaching (23) is unreliable. What makes it so? Where does the unreliability enter in? It will be useful to compare Roxanne's bootstrapping with a procedure that really would establish that a car's gas gauge is reliable. Imagine that Catherine periodically uses a dipstick to measure the level of gas in the tank of her car, and she verifies that the reading of the gauge matches that of the dipstick. Each time she does this, she comes to know an instance of:

---

[29] This is the well-known "generality problem," which itself raises difficulties for reliabilism. See Richard Feldman, "Reliability and Justification," *The Monist*, LXVII (1985): 159-74; and Feldman and Earl Conee, "The Generality Problem for Reliabilism," *Philosophical Studies*, XCIII (1997): 1-29. In presenting my own objections to reliabilism here, I have tried my best to steer clear of problems of this sort, allowing the reliability of the underlying process to be a matter of stipulation. But at this point such complications may be unavoidable.

(21*) On occasion *t*, the gauge is reading accurately.

After enough repeated successes, we can agree, Catherine may know by induction that the gauge is accurate all the time, and she can properly reach the yet stronger (modalized) conclusion that the gauge is reliable. The crucial point is that this impeccable way of finding out that a gas gauge is reliable is identical to Roxanne's original procedure from a certain point onward. Both Catherine and Roxanne infer that their gauges are reliable from beliefs of the form given by (21*). So, it seems that any lapse in reliability on Roxanne's part must occur in the way that she arrives at such beliefs.

In each case, Roxanne deduces an instance of (21*) from a belief of the form:

(20*) On occasion *t*, the gauge reads '*X*' and *X*.

As I said earlier, it would be hard to deny that Roxanne's beliefs about how the gas gauge reads are reliably formed. And, by hypothesis, Roxanne's beliefs about how much gas is in her car's tank are formed by a reliable process. So, the reliabilist seems committed to the claim that when Roxanne comes to believe instances of (20*), those beliefs result from reliable processes. Given what has already been said, the only point left at which some defect can enter into Roxanne's procedure has to be in her transition from instances of (20*) to belief in corresponding instances of (21*). But, since any instance of (20*) entails an instance of (21*), that instance of (21*) must be true if the corresponding instance of (20*) is! In that sense, there is no more risk of error with respect to the instances of (21*) than there is with respect to the instances of (20*). So, it is opaque how proceeding from instances of (20*) to corresponding instances of (21*) turns a reliable process into an unreliable one.[30]

---

[30] Williams insists that reliability cannot be lost through deduction (*op. cit.*, p. 328), leaving him little recourse at this point. Now, the thought may arise again that it is possible for processes to be reliable severally, but not in concatenation. One might then say that the process that led to Roxanne's belief in (20) is reliable, and that deduction is reliable, but that the combination of the two need not count as reliable. Goldman provides good reasons, however, for distinguishing processes that are reliable simpliciter (more precisely, "belief-independent J-rules") from those which are conditionally reliable ("belief-dependent J-rules"). The former generate (largely) true beliefs given nondoxastic inputs. The latter take beliefs as input, and will generate (largely) true beliefs, if their inputs are true. It seems to me that genuine deduction would be a paradigm of a conditionally reliable process, notwithstanding the logical errors made by subjects in certain experimental situa-

Let us also waive *this* objection for the time being, and permit the proponent of NR to take refuge in the claim that bootstrapping is an unreliable process. Thus, her position will be that you cannot use the beliefs generated by some process to establish the accuracy or the reliability of that very process. Unfortunately for the reliabilist, such a response to the gas-gauge case disrupts the full-blooded, noncontextualist antiskepticism that reliabilists often regard as a great virtue of their position. For, while bootstrapping may seem absurd when it involves someone's gas gauge, it has been regarded as a sound response to Cartesian skepticism.[31]

This answer to skepticism is basically G. E. Moore's vintage commonsense response, transposed and fortified by the addition of some early twenty-first-century reliabilism. It goes as follows. You know you have a hand. Your belief that you have a hand was formed by perception, which is a reliable process. Hence, that belief counts as knowledge. Now, the proposition that you have a hand entails that you are not a brain in a vat, deceived into believing falsely that you have a hand. Hence, you can deduce that you are not a brain in a vat from the proposition that you have a hand. Your belief that you are not a brain in a vat qualifies as knowledge, because it, too, is formed by a reliable process. At least, that will be so if deduction from a belief arrived at by a reliable process itself counts as a reliable process.

The correspondence between this neo-Moorean argument and the rejected line of argument in the gas-gauge case emerges clearly if the former is reformulated slightly:

(A) You know you have a hand.
(B) You know that it appears to you as though you have a hand.

---

tions. Goldman allows, as it seems he should, that a belief is reliably formed if it is the output of a conditionally reliable process whose inputs were the output of a reliable process or processes (*Epistemology and Cognition*, especially p. 83). Some care may be necessary here, but I think a reliable process theory must include some provision such as the one Goldman makes. Consequently, I have difficulty envisioning how the reliabilist can legitimately escape the difficulty raised in the text.

[31] For an outline and discussion of reliabilism as an answer to skepticism, see Goldman, *Epistemology and Cognition*, especially pp. 55-57 (though he couches that discussion in terms of a reliabilist theory of justification). The bootstrapping aspect of such a response to skepticism emerges clearly in Alston, pp. 12-17, and in Williams, pp. 327ff. Alston endorses it reluctantly and with qualifications (pp. 138-40). Williams mistakenly believes that his own view avoids the kind of criticism I have raised. It is also noteworthy that reliabilists have embraced bootstrapping in mounting an inductive defense of induction. See, for example, James Van Cleve, "Reliabilism, Justification, and the Problem of Induction," in French, Uehling, and Wettstein, eds., *Midwest Studies in Philosophy*, Volume IX (Minneapolis: Minnesota UP, 1984), pp. 555-67; and David Papineau, "Reliabilism, Induction, and Scepticism," *The Philosophical Quarterly*, XLII (1992): 1-20.

(B) holds, because we may assume that your beliefs about how things appear to you are generated by a reliable process. Continuing:

(C) Therefore, you know that your appearance of having a hand is veridical.
(D) Therefore, you know you are not a deceived brain in a vat.

For comparison:

(A′) Roxanne knows that her gas tank is full.
(B′) Roxanne knows that the gas gauge reads 'F'.
(C′) Therefore, Roxanne knows that, on this occasion, her gas gauge is reading accurately.

Thus, the reliabilist version of Moore's refutation of skepticism sanctions the same kind of inference that created problems in connection with the gas-gauge case. Suppose we grant that the reliabilist is able to evade those problems by denying, either for principled reasons or by fiat, that bootstrapping can lead to knowledge. The upshot will be that reliabilism cannot provide a satisfactory response to skepticism along the lines just laid out.[32]

Let me return to the status of NR in general. I have taxed that view for allowing bootstrapping to count as a source of knowledge, or for not being able to explain what is wrong with it. In the example given above, the crux of the matter came down to the status of beliefs of the form:

(20*) On occasion t, the gauge reads 'X' and X.

and:

(21*) On occasion t, the gauge is reading accurately.

It looked as though the reliabilist was committed to saying that Roxanne knows instances of (20*), but does not know corresponding instances of (21*), even though she deduces the second from the first. Now, contextualism about knowledge might appear to offer the reliabilist a way out of this uncomfortable situation. The contextualist will say that our standards for knowledge rise as we think through the example. That is, the neighborhood N which determines the reliability of processes expands. At first, when (20*) is under consideration, worlds in which the gas gauge fails are excluded from *N*. The gas gauge counts as reliable, and Roxanne is correctly judged to know both conjuncts of (20*), and (20*) itself. But overt consideration of (21*) brings questions about the accuracy of the gauge into the picture, and worlds in which the gauge is wrong come to be included in *N*. In this setting, the gauge no longer counts as reliable, so Roxanne does not know that her gas tank is full. Hence, she does not know one of the conjuncts of (20*), and she does not know (20*) as a whole. Nor does she know (21*), which she derived from (20*). Contextualism relieves the reliabilist's embarrassment of having to explain how reliability is lost in the course of Roxanne's deducing a proposition from another one that entails it.

---

[32] One might think that the reliabilist can simply stand on the claim that we have perceptual knowledge of the external world, insofar as our perceptual beliefs are generated by a reliable process. We cannot, on this more minimal view, move on to knowledge that our experiences are veridical; that further step would allow for bootstrapping, which we have said is unreliable. One difficulty, among others, with such a position is that it is inconsistent with the closure principle for knowledge.

I do not propose to enter into a full-blown discussion of contextualism here, but I shall make two brief points. First, if the standards of reliability and knowledge can be elevated in something like the manner just described, then we almost never know explicitly that we know anything. We also lose some seemingly impeccable first-order knowledge. To see this, recall the example in which you believe that Omar has new shoes. Say that you reflect upon your epistemic situation, and you believe that your belief that Omar has a new pair of shoes is the outcome of a reliable process. According to the contextualist revision of NR, considering the reliability of a belief-forming process expands the neighborhood *N* so that the process no longer counts as reliable. It follows that once you reflect upon the fact that your belief-forming process is reliable, you no longer know that Omar has a new pair of shoes. A fortiori, you do not know that you know that Omar has new shoes. Recourse to contextualism of this sort would make both first- and higher-order knowledge very fragile, unable to survive virtually any reflection at all. Such a result is hardly acceptable.[33]

My second thought is that adding a contextualist dimension to NR may be going to great lengths to save a view that does not want to be saved. I say this because a consequence of the contextualist maneuver is that, if it is possible to maintain low epistemic standards while the entire bootstrapping procedure is carried out, that procedure would be successful. Roxanne could convert knowledge of how much fuel is in the tank of her car into knowledge that her gas gauge is reliable.

---

[33] To be sure, the contextualist scheme I have described is quite crude, and one might hope to find more satisfactory accounts of when and how epistemic standards get raised. But I am not aware of any worked-out proposal that would serve the reliabilist's purposes any better; see my "The New Relevant Alternatives Theory," *Philosophical Perspectives*, XIII (1999): 155-80, especially section 3.

But, for the reliabilist, knowledge of how much fuel is in the tank is knowledge of one empirical fact, and knowledge that the gauge is reliable is knowledge of another, independent empirical fact. One should not be able to transmute the first into the second, as bootstrapping would allow one to do.[34] To put the same point another way: suppose a change in epistemic standards of the sort envisioned does thwart or suppress the bootstrapping procedure. Still, the standards' tendency to change would not eliminate the possibility of bootstrapping in principle or somehow make bootstrapping intrinsically more acceptable. To that extent, whatever relief contextualism provides will strike the reliabilist as cosmetic, not real.

IV. SOME CONCLUSIONS

CR and NR, as I understand them, are meant to be general theories of what knowledge is. I have been arguing that both views encounter substantial difficulties when they have to deal with higher-level knowledge. CR allows:

(10) You know that Omar has new shoes.

but not the higher-level:

(11) You know that your belief that Omar has new shoes is true (not false).

NR got into trouble with the gas-gauge case. A proponent of NR has to say that Roxanne knows:

(20) On this occasion, the gauge says 'F' and *F*.

But if he goes on to grant that she knows:

(21) On this occasion, the gauge is reading accurately.

it appears that she can proceed to knowledge of:

(23) The gauge is reliable.

and beyond. This result is unacceptable, yet it is not clear how NR can avoid it.

What lessons should we draw? First, these phenomena provide some reason—not a conclusive reason, but some reason—to endorse the traditional view that knowledge requires justification. The traditional view handles the problem cases with relative ease. First, it offers a natural explanation of why (11) follows so closely on (10). Good evidence that

---

[34] I suppose Roxanne also has reliable beliefs, and knowledge, about the outputs of the gauge. What she still lacks is any access to whether the gauge itself is accurate or reliable.

Omar has new shoes is good evidence that a belief that Omar has new shoes is true, not false. If you have such evidence (and you believe that Omar has new shoes, and your belief is true, and there are no Gettier traps), then you are in a good position to have the knowledge reported in (10). Moreover, given that you have this evidence, you are also in a good position to have the higher-level knowledge reported in (11), so long as you recognize that you believe that Omar has new shoes, and you believe that your belief is true, not false.

Bringing justification into the picture also helps to explain what is wrong with Roxanne's bootstrapping. Given the way the example was framed, Roxanne has no justification for (23), nor, I think, for (22) or (21). If justification is necessary for knowledge, she does not know any of these, which seems correct. What about (20)? Since (20) directly entails (21), Roxanne would be justified in believing (21) if she is justified in believing (20). If, as I claim, she is not justified in believing (21), then she is not justified in believing (20), either. That, too, seems plausible. If Roxanne has no independent reason to believe that the position of the needle on the gauge is reliably correlated with how much gas is in the tank, I cannot see that she is justified in believing that her tank is full when she looks at the gauge. So the traditionalist can say that Roxanne does not know (20) or (21), since she lacks justification for both. There are no bootstraps for her to pull on.

Thus, re-instating a justification requirement for knowledge does make it easier to explain what is going on in the gas-gauge case. An objection is bound to arise, however. I have as much as said that for Roxanne to know how much gas she has by reading the gauge, she needs to have justification for thinking that the gauge is reliable. This claim seems to invite a version of the regress problem discussed in section I. Suppose that, in general, justification for believing that X always requires justification for the belief that the process by which you came to believe that X is reliable. It apparently follows that justification for any belief will require an endless hierarchy of further justified beliefs of level $(N + 1)$ as to the reliability of the way one's belief at level N was formed.[35] But to say that justification *sometimes* requires reason to believe that one's belief-forming process is reliable, as it does in Roxanne's situation, does not imply that justification *always* requires such higher-level support or supplementation. It is this second, stronger claim that creates the regress, and one can refuse to agree to it. Note, though, that one can maintain that knowledge always requires justification, without maintaining that justification always requires *further* justification for the higher-level belief that the process that led to the formation of one's initial belief was reliable.[36]

---

[35] Some coherentists cheerfully accept this consequence.

In sum, reliabilism encounters serious difficulties in accounting for reflective knowledge, difficulties that might well be avoided by a justificationist approach to knowledge. The question now arises whether the weakness in reliabilism that has come to light is confined to its treatment of reflective knowledge, or whether there is some deeper, more far-reaching problem with reliabilism which vividly shows itself when reflective knowledge is under consideration. Correlatively, one might ask whether the only essential work the justification condition does is to capture the relations between lower- and higher-level knowledge.

My own view, which I shall state but not argue for, is that the inability of reliabilism to account for reflective knowledge has its roots in a more basic deficiency. I take it that knowledge is a kind of human success. It is something we accomplish; we have to do our part. The justification requirement is an attempt to formulate what that part is. Reliabilism either loses sight of this aspect of knowledge, or tries, in effect, to substitute some kind of de facto alignment between belief and the world for the knower's own contribution.[37] As a consequence, reliabilist theories go wrong in their treatment of various cases, and these include the examples deployed above. Reliabilism has trouble accounting for reflective knowledge because it is defective as an account of knowledge in general. A consideration of reflective knowledge shows that knowledge in general is neither reliably true belief nor belief that results from a reliable process.

JONATHAN VOGEL

Amherst College

---

[36] For such a view, see Alston, "What's Wrong with Immediate Knowledge?" reprinted in his *Epistemic Justification: Essays in the Theory of Knowledge* (Ithaca: Cornell, 1989), pp. 57-78. Another response would be to deny that the pertinent beliefs fall into the hierarchy of levels described in the text, setting off the regress. Perhaps first-order beliefs about how various mechanisms perform can justify beliefs about the reliability of those mechanisms, and about the reliability of belief-forming procedures that make use of those mechanisms. A view something like this was developed by Wilfrid Sellars, "Empiricism and the Philosophy of Mind," reprinted in *Science, Perception and Reality* (London: Routledge and Kegan Paul, 1963), pp. 127-96.

[37] So-called "virtue epistemology" seeks to do justice to the contribution of the knower while retaining many of the central features of reliabilism. My impression is that the reliabilist component of such accounts prevents them from giving a proper treatment of the way a knower's *evidence* figures in knowledge, but I shall not pursue that issue here.

A NOTE ON WITTGENSTEIN'S "NOTORIOUS PARAGRAPH" ABOUT THE GÖDEL THEOREM*

> I imagine someone asking my advice; he says: "I have constructed a proposition (I will use 'P' to designate it) in Russell's symbolism, and by means of certain definitions and transformations it can be so interpreted that it says: 'P is not provable in Russell's system'. Must I not say that this proposition on the one hand is true, and on the other hand is unprovable? For suppose it were false; then it is true that it is provable. And that surely cannot be! And if it is proved, then it is proved that it is not provable. Thus it can only be true, but unprovable."
>
> Just as we ask, "'Provable' in what system?," so we must also ask, "'True' in what system?" "True in Russell's system" means, as was said, proved in Russell's system, and "false in Russell's system" means the opposite has been proved in Russell's system.—Now what does your "suppose it is false" mean? *In the Russell sense* it means, "suppose the opposite is proved in Russell's system"; *if that is your assumption* you will now presumably give up the interpretation that it is unprovable. And by "this interpretation" I understand the translation into this English sentence. —If you assume that the proposition is provable in Russell's system, that means it is true *in the Russell sense,* and the interpretation "P is not provable" again has to be given up. If you assume that the proposition is true in the Russell sense, *the same* thing follows. Further: if the proposition is supposed to be false in some other than the Russell sense, then it does not contradict this for it to be proved in Russell's system. (What is called "losing" in chess may constitute winning in another game.)[1]

We believe that this "notorious paragraph"[2] contains a philosophical claim of great interest which has been almost entirely missed in the brouhaha about whether Ludwig Wittgenstein "misunderstood" Gödel's first Incompleteness Theorem. Our purpose here is to detach that claim, so to speak, from that disputed question (although the fact that Wittgenstein's critics seem to have missed it must surely be *relevant* to the dispute).

---

*Floyd is grateful to Jan Harald Alnes, Akihiro Kanamori, and Rohit Parikh for conversations on points related to the issues discussed in this paper.

[1] Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics,* G.H. von Wright, R. Rhees, and G.E.M. Anscombe, eds.; Anscombe, trans. (Cambridge: MIT, 1978, revised edition), I, Appendix III, §8.

[2] So-called by Floyd in her "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth," *Philosophia Mathematica* (forthcoming). Floyd gives a detailed reading of this paragraph in "On Saying What You Really Want to Say: Wittgenstein, Gödel and the Trisection of the Angle," in Jaakko Hintikka, ed., *From Dedekind to Gödel: Essays on the Development of the Foundations of Mathematics* (Boston: Kluwer, 1995), pp. 373-425.

0022-362X/00/9711/624-32

The claim is simply this: if one assumes (and, a fortiori if one actually finds out) that $\neg P$ is provable in Russell's system one should (or, as Wittgenstein actually writes, one "will now presumably") give up the "translation" of P by the English sentence 'P is not provable'.

To see that Wittgenstein is on to something here, let us imagine that a proof of $\neg P$ has actually been discovered. Assume, for the time being, that Russell's system (henceforth PM) has not actually turned out to be *inconsistent*, however. Then, by the first Incompleteness Theorem, we know that PM is ω-inconsistent. But what does ω-inconsistency show? ω-inconsistency shows that a system *has no model in which the predicate we have been interpreting as* 'x is a natural number' *possesses an extension that is isomorphic to the natural numbers.*

But why did we "translate" P as 'P is not provable in PM'? Well, P has the form: $\neg(\exists x)(\mathrm{NaturalNo.}(x).\mathrm{Proof}(x,t))$, where '$t$' abbreviates a numerical expression whose value calculates out to be the Gödel number of P itself, 'Proof' abbreviates a predicate that is supposed to define an effectively calculable relation that holds between two natural numbers n,m just in case n is the Gödel number of a proof whose last line is the formula with Gödel number m, and 'NaturalNo.(x)' is the predicate of PM we interpret as 'x is a natural number'.

But in discovering that PM is ω-inconsistent we have discovered that:

(1) 'NaturalNo.(x)' *cannot* be so interpreted. In all admissible interpretations of PM (all interpretations that fit at least one *model* of PM), there are entities that are not *natural numbers* (and, a fortiori, not Gödel numbers of proofs).

(2) Those predicates of PM (for example, 'Proof(x,t)') whose extensions are provably infinite, and which we believed to be infinite subsets of N (the set of all natural numbers), do not have such extensions in any model. Instead, they have extensions that invariably also contain elements that are not natural numbers.

In short, the "translation" of P as 'P is not provable in PM' is untenable in this case—just as Wittgenstein asserted! This does not, however, affect the correctness of Gödel's proof, for nothing in that proof turns on any such translation into ordinary prose.[3]

---

[3] In introductory remarks to his 1931 paper, Gödel did sketch an argument using intuitive versions of the notions of *truth* and *proof.* But he was quite explicit that this heuristic sketch plays no essential role in the proof he gives with "full precision" in the body of the paper; see "On Formally Undecidable Propositions of Principia Mathematica and Related Systems I," *Kurt Gödel: Collected Works, Volume I,* Solomon Feferman et alia, eds. (New York: Oxford, 1986), pp. 147-51, especially p. 147 note 6, p. 151 paragraph 2, p. 173 note 41.

Wittgenstein's aim is not to refute the Gödel theorem but to "by-pass" it.[4]

In addition, we may point out that, if PM is actually inconsistent, and not merely ω-inconsistent, then it has no "admissible interpretations"—which is not to deny that in various contexts, and for various reasons, we may want to correlate its sentences with sentences in English.

I. BUT SURELY WITTGENSTEIN COULDN'T HAVE KNOWN ABOUT THAT STUFF!

But is it believable that Wittgenstein could have known about the "numerical insegregativity"[5] of ω-inconsistent systems? Are we not being overly charitable in giving this as Wittgenstein's reason for saying that, if one assumes that $\neg P$ is provable in Russell's system, one "will now presumably give up the interpretation" of P by the English sentence 'P is unprovable'?

The answer is that we have testimony (and not from a particularly sympathetic source!) that Wittgenstein thought about what are now called "nonstandard" models of the natural numbers, and connected them with the Gödel theorem. In 1957, R.L. Goodstein[6] wrote:

> Wittgenstein with remarkable insight said in the early thirties that Gödel's results showed that the notion of a finite cardinal could not be expressed in an axiomatic system and that formal number variables must necessarily take values other than natural numbers; a view which, following Skolem's 1934 publication, of which Wittgenstein was unaware, is now generally accepted (*ibid.*, p. 551).

And in 1972, Goodstein[7] wrote:

> I do not think Wittgenstein heard of Gödel's discovery before 1935; on hearing about it his immediate reaction, with I think truly remarkable insight, was to observe that it showed that the formalization of arithmetic with mathematical induction and the substitution of numerals for vari-

---

[4] See *Remarks on the Foundations of Mathematics*, VII §19: "My task is, not to talk about (e.g.) Gödel's proof, but to by-pass it."

[5] A term introduced by Quine to refer to the model-theoretic fact that ω-inconsistent systems have no model in which the "integers" of the model are (isomorphic to) the natural numbers. See his "On ω-inconsistency and the So-called Axiom of Infinity," in *Selected Logic Papers* (Cambridge: Harvard, 1995, enlarged edition), pp. 114-20; compare *Set Theory and Its Logic* (Cambridge: Harvard, 1969, revised edition), pp. 305-06.

[6] "Critical Notice of Remarks on the Foundations of Mathematics," *Mind*, LXVI (1957): 549-53.

[7] "Wittgenstein's Philosophy of Mathematics," in A. Ambrose and M. Lazerowitz, eds., *Ludwig Wittgenstein: Philosophy and Language* (London: Allen and Unwin, 1972), pp. 271-86.

> ables fails to capture the concept of natural number, and the variables must admit values which are not natural numbers. For if, in a system *A*, all the sentences G(n) with n a natural number are provable, but the universal sentence (n)G(n) is not, then there must be an interpretation of *A* in which n takes values other than natural numbers for which G(n) is not true (in fact in 1934, Th. Skolem had shown that this was the case, independently of Gödel's work) (*ibid.*, p. 279).

Interestingly, Goodstein entirely missed the connection between this "remarkable insight" and the claim we are discussing—which may not only be taken to be contained in Appendix I, §8 of *Remarks on the Foundations of Mathematics*, but also as constituting virtually the whole of §10 of the same Appendix![8] But, as is well known, the remarks on Gödel's theorem were written as notes for Wittgenstein himself, and there was no reason for their author to state explicitly everything that he knew in connection with them. The fact is that Wittgenstein did understand—"with remarkable insight"—that "variables must necessarily take on values other than the natural numbers." (Indeed, this must be the case whether PM is ω-inconsistent or not; the point about ω-inconsistency is that, if PM is ω-inconsistent, then there is *no* interpretation under which PM's theorems come out true in which the formal number variables take *only* the natural numbers as values.) We know this not only from Goodstein, but also from remarks made by Wittgenstein's student Alister Watson[9] in a 1938 paper in *Mind*. Watson explicitly states that his interpretation of the Gödel incompleteness result "owes much to lengthy discussions with a number of people, especially Mr. Turing and Dr. Wittgenstein" (*ibid.*, p. 445). He then gives an argument very similar to the one we sketched above about the effect of ω-inconsistency on the power of any recursively axiomatized system of arithmetic to express the notion of natural number. He does this, he says, in order to show that the intuitive argument about the true but unprovable proposition P "obscures rather than illuminates the point" of Gödel's theorem.[10] This, as we have just argued, may be seen to have been Wittgenstein's point as well in the "notorious paragraph."

---

[8] *Remarks on the Foundations of Mathematics*, I, Appendix III, §10: " 'But surely P cannot be provable, for, supposing it were proved, then the proposition that it is not provable would be proved'. But if this were now proved, or if I believed—perhaps through an error—that I had proved it, why should I not let the proof stand and say I must withdraw my interpretation '*unprovable*'?"

Compare Goodstein's dismissal of Wittgenstein's remarks on Gödel, "Critical Notice of Remarks on the Foundations of Mathematics," p. 551, and "Wittgenstein's Philosophy of Mathematics," p. 279.

[9] "Mathematics and Its Foundations," *Mind*, XLVII (1938): 440-51. According to Andrew Hodges, Turing's biographer, discussions were held between Watson, Turing, and Wittgenstein in the summer of 1937, when Turing returned for a season to Cambridge between his years at Princeton—*Alan Turing: The Enigma* (New York: Touchstone, 1983), pp. 109, 136. The "notorious paragraph" of *Remarks on the Foundations of Mathematics*, I, Appendix III, §8 was penned on September 23, 1937, when Wittgenstein was in Norway (see the Wittgenstein papers, CD Rom, Oxford University Press and the University of Bergen, 1998, Item 118 (Band XIV), pp. 106ff.). According to G.H. von Wright (correspondence with Floyd) Wittgenstein praised Watson's paper very highly on more than one occasion—something Wittgenstein did not very often do with colleagues' work.

II. BUT EVEN SO...

Even though, as we explained, our purpose here is to examine the valuable and still relatively unappreciated point that Wittgenstein makes in §8 (and also in §10) of Appendix I of *Remarks on the Foundations of Mathematics*—that is, that the "translation" of the famous Gödel sentence P as 'P is unprovable in PM' is not cast in stone, but is something that we have to give up in certain contexts—rather than to discuss in detail the dispute(s) about (1) whether Wittgenstein "understood" the Gödel theorem, and (2) whether his philosophical remarks about it have any value, we would be remiss if we did not at least comment on the question of the bearing of the unappreciated point on at least the second of those disputes. (The reader who has followed us this far should have no doubts about whether Wittgenstein understood the Gödel theorem, and hence about the position to take with respect to the first dispute!)

---

[10] Watson, "Mathematics and Its Foundations," pp. 446-47:

"If we assume for the moment that this axiomatic system is indeed a good basis for arithmetic, we shall have to conclude that the formula is not provable, and therefore, since this is just what it says, that it is true. For if it were provable, it would be false, and the system would be incorrect.

This method of putting the argument, however, obscures rather than illuminates the point. Suppose we assume the falsity of the formula, we cannot, of course, derive a contradiction, for this would amount to a proof of the formula. Instead, we reach the following peculiar situation, which is called by Gödel an ω-contradiction (ω is the "ordinal number" of a sequence). We find that there is a function of a cardinal variable, say $\mathrm{f(n)}$, such that (all on the basis of the falsehood of Gödel's formula) $\mathrm{(n)f(n)}$ can be disproved, and yet we can convince ourselves that we can prove in turn $\mathrm{f(0),f(1),f(2)}$ and so on. In other words, we apply mathematical induction to the proofs of the system, and obtain $\mathrm{f(0)}$, and from a proof of $\mathrm{f(n)}$ for any particular value of n, a proof for $\mathrm{n+1}$.

Why should we object to an ω-contradiction? Why should we not still say that Gödel's formula may be false? The answer is that if we do this we shall feel compelled to say that the cardinal numbers cannot be *all* the values of the variable n, if $\mathrm{f(n)}$ can be true for each particular value of n, and yet $\mathrm{(n)f(n)}$ be false. ...

Thus the notion of a cardinal variable, i.e. of a number in the everyday sense, is something that cannot be completely expressed in the axiomatic system."

We note that although Watson is right that, if we assume the falsity of Gödel's formula—that is, if we assume "$(\exists x)(\text{NaturalNo.}(x).\text{Proof}(x,t))$"—then $(n)f(n)$ can be disproved (taking '$f(n)$' to be "$\neg(\text{NaturalNo.}(n).\text{Proof}(n,t))$" and yet "we can prove in turn $f(0), f(1), f(2)$ and so on," the usual proof of this ω-inconsistency does not proceed by "applying mathematical induction to the proofs of the system." Rather, one employs the fact that, if PM is inconsistent that it is (trivially) ω-inconsistent, and if it is consistent, then by the already proven facts that (i) PM would be inconsistent if any natural number n were the Gödel number of a proof of the formula with Gödel number 't', and (ii) '$\text{Proof}(n,t)$' represents the recursive relation "n is a proof of the formula with Gödel number t" and "NaturalNo.(0)," "NaturalNo.(1)," "NaturalNo.(2)"...are all (trivially) provable, it follows that for each numeral 'n', "$\neg(\text{NaturalNo.}(n).\text{Proof}(n,t))$" is provable in PM, that is, that there are proofs of $f(0), f(1), f(2)$.... We see no way to use mathematical induction in the way Watson suggests. This error is irrelevant, however to the point Watson was making.

The principal source of the question as to the cogency of Wittgenstein's discussion are the words in the notorious §8: "'True in Russell's system' means, *as was said*: proved in Russell's system, and 'false in Russell's system' means: the opposite has been proved in Russell's system" (emphasis added). Is Wittgenstein not aware that arithmetical propositions have a meaning and truth value *independent* of what system they are formalized in? Is he just identifying "truth" and "provability" out of some misguided combination of formalist and constructivist/finitist motives? And so on.

The answer is that one cannot simply ignore the direction (implicit in the words we italicized) to look at what was said *before* about 'true in Russell's system'.

Just one paragraph before (§7), there appear the remarks which clearly set the stage for the notorious remark:

> 7. "But may there not be true propositions which are written in this symbolism, but are not provable in Russell's system?"—'True propositions', hence propositions which are true in another system, i.e. can rightly be asserted in *another* game. Certainly; why should there not be such propositions; or rather: why should not propositions—of physics, e.g.—be written in Russell's symbolism? The question is quite analogous to: Can there be true propositions in the language of Euclid, which are not provable in his system, but are true?—Why, there are even propositions which are provable in Euclid's system, but are *false* in another system. May not triangles be—in another system—similar (*very* similar) which do not have equal angles?—"But that's just a joke! For in that case they are not 'similar' to one another in the same sense!"—Of course not; and a proposition which cannot be proved in Russell's system is "true" or "false" in a different sense from a proposition of *Principia Mathematica.*[11]

If one reads this paragraph with care, one will observe two things: first, that Wittgenstein is telling us that we should look on PM as a "system" in the sense in which a system of non-Euclidean geometry is a "system" of geometry—a sense in which the same sentence (*Satz*) can be true in one system and false in another. And, second, that this paragraph does *not* deny that a proposition that cannot be proved in Russell's system (Wittgenstein obviously means one that cannot be decided, that is, proved or disproved) can, in *some* sense be "true" or "false" (outside the system)—he only asserts that this is a "different sense" from the sense in which it is true or false *as* a "proposition of *Principia Mathematica*." We shall now comment on each of these points in turn.

---

[11] *Remarks on the Foundations of Mathematics*, I, Appendix III, §7.

(1) Wittgenstein's targets in much of Part I of *Remarks on the Foundations of Mathematics* are Gottlob Frege and Bertrand Russell, not as mathematical logicians but as philosophers of mathematics and logic. These philosophers emphatically did not see themselves as providing a mere *notation* into which one could transcribe the propositions that mathematicians actually utter, write, and publish in ordinary "mathematical prose," that is, in English or French or German or.... They saw themselves as providing a freestanding "ideal language" or "concept-language," what W. V. Quine[12] has called a first-grade conceptual scheme, which in some sense *supercedes* ordinary language. Moreover, in providing such a scheme, they saw themselves as providing mathematics with a *foundation*. Ordinary language might be necessary to "lead someone into" the ideal language, but the "elucidations" we offer for this purpose in ordinary language are, so to speak, a ladder that we can throw away. Frege explicitly argues that ordinary language sentences that we use to explain the ideal notation do not and cannot capture the precise content of the ideal notion.[13] It would be utterly foreign to this spirit to explain the truth of a formula of *Principia Mathematica* by merely writing down an English sentence, and saying this is what it means for *P* to be true. To confess that this is what one has to do would be to abandon the claim for the *foundational* status of a system such as *Principia Mathematica* entirely. (So that when Wittgenstein writes in §8: "And by 'this interpretation' *I understand the translation into this English sentence*," he is already denying that this notion of an "interpretation" which can only be indicated in English by helpful "hints"[14]

---

[12] "Speaking of Objects," *Ontological Relativity and Other Essays* (New York: Columbia, 1969), p. 24.

[13] For the case of what Frege takes to be his primitive or undefinable notions (for example, *function, concept*), see his *Collected Papers on Mathematics, Logic, and Philosophy*, Brian McGuinness, ed., Max Black et alia, trans. (New York: Blackwell, 1984), pp. 193-94, 300, 302; *Posthumous Writings*, pp. 207, 214, 235; *Grundgesetze der Arithmetik, I* (Hildesheim: Olms, 1966), Appendix 2, n. 1 (p. 240).

[14] Compare Frege, "On Concept and Object," *Collected Papers*, p. 194.

and which does not in principle require any dependence on informal mathematical language, makes any real sense.)

Instead, Wittgenstein is suggesting, there *is* a sense in which a formalism can be free-standing, but it is not the sense of a *Begriffsschrift* or an ideal language. It can be free-standing regarded simply as a formal system. But then the only sense of "truth" we shall have available is: being a theorem of the system.[15] And in that case, why should it not be the case that one and the same "proposition" should be "true in Russell's system" and false in a *different* system?

Indeed, something of this kind does happen. *If there are only finitely many individuals, then for some natural number x,* $x=Sx$ (imagine this written out in the formal notation) is a theorem of *Principia Mathematica,* while *There are only finitely many individuals and for every natural number x,* $x \neq Sx$ (imagine this likewise written out in formal notation) is a consistent proposition (model theoretically as well as proof theoretically) in Zermelo set theory. Or to change the example, the formula '$(\exists x)(x \in x)$' ("some set belongs to itself") is "true in Quine's system" (*Mathematical Logic*) and "false in Zermelo set theory"!

Today, of course, few if any philosophers think that a formal system provides a foundation for either the content or the truth of mathematical propositions. But one of us (Putnam) remembers a delightful philosophical conversation between C.G. Hempel and one of Hans Reichenbach's graduate students in Reichenbach's living room in 1950, at which the older attitude and the newer attitude memorably clashed. Hempel was defending Quine's skepticism with respect to the analytic-synthetic distinction, and the graduate student said plaintively: "Quine's arguments may show that the analytic-synthetic distinction makes no sense in natural language. But why doesn't it make clear sense in a *formalized* language?"; and Hempel replied: "Every formalized language is ultimately explained in some natural language. The disease [Hempel meant the unclarity of the analytic-synthetic distinction] is *hereditary.*" Here, Hempel—like Wittgenstein in *Remarks on the Foundations of Mathematics*—was denying that a formal system *could* provide us with a standard of truth or clarity that is, in principle, inaccessible to a natural language.

---

[15] After the appearance of semantics with Alfred Tarski, one can add another sense: holding in all models of the system. If we take this to mean all Henkin models, then even in the case of a higher-order language like PM, this will coincide with being a theorem. It is a further question, however, whether Tarski's model-theoretical account of truth definitions for formalized languages exhibits a correct analysis of our intuitive notion of *truth.* On the claim that Tarski analyzed the *concept* of truth, see Floyd, "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth"; and also Putnam, "A Comparison of Something with Something Else," in his *Words and Life* (Cambridge: Harvard, 1994), pp. 330-50. So far as we know, Wittgenstein never mentioned Tarski's work on truth definitions.

(2) What if someone were to have said to Wittgenstein: "When I say that *P* is true in Russell's system, what I mean is simply that its translation into English—any one of its mathematically equivalent translations, including '*P* is unprovable in PM'—is true?" We believe that Wittgenstein would have pointed out that the notion of *truth* is eliminable here.[16] To understand '*P* is true in PM' as meaning "The English sentence '*P* is unprovable in PM' is true" (or more colloquially, as meaning "It is true that *P* is unprovable in PM"), would amount (as we see by "disquoting") just to understanding '*P* is true in PM' as short for '*P* is unprovable in PM', or understanding *P* itself (in PM) as short for '*P* is unprovable in PM'. In short this just is to accept what Wittgenstein calls "the translation into this English sentence." And this is something, we have shown, Wittgenstein *did* discuss in the "notorious paragraph"—"with remarkable insight."

That the Gödel theorem *shows* that (1) there is a well-defined notion of "mathematical truth" applicable to every formula of PM; and (2) that, if PM is consistent, then some "mathematical truths" in *that* sense are undecidable in PM, is *not* a mathematical result but a metaphysical claim. But that if *P* is provable in PM then PM is inconsistent and if $\neg P$ is provable in PM then PM is ω-inconsistent is precisely the mathematical claim that Gödel proved. What Wittgenstein is criticizing is the philosophical naiveté involved in confusing the two, or thinking that the former follows from the latter. But not because Wittgenstein wants simply to *deny* the metaphysical claim; rather, he wants us to see how little sense we have succeeded in giving it.

JULIET FLOYD

Boston University

HILARY PUTNAM

Harvard University

---

[16] For a discussion of the senses in which Wittgenstein *was* and *was not* a "disquotationalist," cf. Putnam, Lecture III in Part I (The Dewey Lectures) of his *The Threefold Cord: Mind, Body, and World* (New York: Columbia, 1999).

NEW BOOKS: PAPERBACK REPRINTS

ANSCOMBE, G.E.M.: *Intention.* Cambridge: Harvard, 1957/2000. ix, 94 p. Paper $14.95.

APPLBAUM, ARTHUR ISAK: *Ethics for Adversaries: The Morality of Roles in Public and Professional Life.* Princeton: University Press, 1999/2000. xii, 273 p. Paper $16.95.

BALSLEV, ANINDITA NIYOGI: *Cultural Otherness: Correspondence with Richard Rorty.* Atlanta: Scholars Press, 1991/2000. 125 p. Paper $16.95.

BARWISE, JON, and JOHN PERRY: *Situations and Attitudes.* New York: Cambridge, 1983/1999. lxxxviii, 352 p. Paper $22.95.

BERMÚDEZ, JOSÉ LUIS: *The Paradox of Self-Consciousness.* Cambridge: MIT, 1998/2000. 338 p. Paper $18.00.

CASSAM, QUASSIM: *Self and World.* New York: Oxford, 1997/1999. viii, 208 p. Paper $19.95.

FODOR, JERRY: *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind.* Cambridge: MIT, 1998/2000. x, 219 p. Paper $16.95.

GALLAGHER, SHAUN, and JONATHAN SHEAR, eds.: *Models of the Self.* Bowling Green, OH: Imprint Academic, 1999/2000. xx, 524 p. Paper $29.50.

GRAY, JOHN: *Voltaire.* New York, Routledge, 1997/1999. 54 p. Paper $6.00.

HEGEL, G. W. F.: *Natural Law: The Scientific Ways of Treating Natural Law, Its Place in Moral Philosophy, and Its Relation to the Positive Sciences of Law.* T. M. KNOX, trans. Philadelphia: Pennsylvania UP, 1975/2000. 137 p. Paper $14.95.

KENNY, ANTHONY: *Frege: An Introduction to the Founder of Modern Analytic Philosophy.* Malden, MA: Blackwell, 1995/2000. xi, 223 p. Paper $24.95.

KIM, JAEGWON: *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation.* Cambridge: MIT, 1998/2000. viii, 146 p. Paper $15.00.

LOVIBOND, SABINA, and S. G. WILLIAMS, eds.: *Identity, Truth, and Value: Essays for David Wiggins.* Malden, MA: Blackwell, 1996/2000. ix, 306 p. Paper $29.95.

MASON, RICHARD: *The God of Spinoza: A Philosophical Study.* New York: Cambridge, 1997/1999. xiv, 272 p. Paper $21.95.

MCCABE, MARY MARGARET: *Plato's Individuals.* Princeton: University Press, 1994/1999. xiii, 339. Paper $18.95.

MILLGRAM, ELIJAH: *Practical Induction.* Cambridge: Harvard, 1997/1999. vii, 184 p. Paper $17.95.

MONK, RAY: *Russell.* New York: Routledge, 1997/1999. 58 p. Paper $6.00.

MOSER, PAUL K.: *Philosophy after Objectivity: Making Sense in Perspective.* New York: Oxford, 1993/1999. xii, 267 p. Paper $19.95.

PIEPER, JOSEF: *Death and Immortality.* RICHARD WINSTON and CLARA WINSTON, trans. South Bend, IN: St. Augustine's, 1969/2000. 134 p. Paper $11.00.

PITKIN, HANNA FENICHEL: *Fortune Is a Woman: Gender and Politics in the Thought of Niccolò Machiavelli.* Chicago: University Press, 1984/1999. ix, 374 p. Paper $16.00.

RAZ, JOSEPH: *Practical Reason and Norms.* New York: Oxford, 1975/1999. 220 p. Paper $29.95.

RESCHER, NICHOLAS: *Philosophical Standardism: An Empirical Approach to Philosophical Methodology.* Pittsburgh: University Press, 1994/2000. ix, 214 p. Paper $19.95.

ROSENBLUM, NANCY L.: *Membership and Morals: The Personal Uses of Pluralism in America.* Princeton: University Press, 1998/2000. vii, 438 p. Paper $18.95.

SELIGMAN, ADAM B.: *The Problem of Trust.* Princeton: University Press, 1997/2000. 224 p. Paper $16.95.

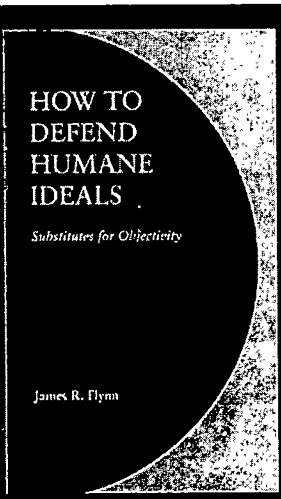
VARDY, PETER: *What Is Truth?* Sydney, Australia: UNSW Press, 1999/2000. x, 202 p. Paper $24.95.

WERTHEIMER, ALAN: *Exploitation.* Princeton: University Press, 1996/1999. xiv, 316 p. Paper $19.95.

WIGGINS, DAVID: *Needs, Values, Truth: Essays in the Philosophy of Value.* New York: Oxford, 1998/2000. xii, 398 p. Paper $19.95.

WILLIAMS, BERNARD: *Plato.* New York: Routledge, 1997/1999. 57 p. Paper $6.00.

WOLLHEIM, RICHARD: *The Thread of Life.* New Haven: Yale, 1984/1999. xv, 288 p. Paper $18.00.



UNIVERSITY OF NEBRASKA PRESS
Publishers of Bison Books
www.nebraskapress.unl.edu
800-755-1105

HOW TO DEFEND HUMANE IDEALS
Substitutes for Objectivity
JAMES R. FLYNN
"Time and again, Flynn has spurred debate and opened up entirely new fields of scholarship. This book is an ambitious attempt to integrate philosophy, psychology, and sociology to provide an evidential case for humane ideals."—Stephen J. Ceci, editor of *The Nature Nurture Debate: The Essential Readings.* "Flynn shows that a humane and progressive liberalism can produce powerful arguments in the social and political realm, while abandoning what he regards as discredited and unhelpful notions of ethical proof and objectivity. There is nothing quite like this in literature—nothing that matches the power of Flynn's meta-ethical discussion and sustained first-level arguments. The chapters on the dilemmas and difficulties in modern social policy on race and class are magnificent."
—Jeremy Waldron, author of *Social Justice.*
$40 cloth

Best in Philosophy from Cambridge

Cambridge Companion to German Idealism
Karl Ameriks, *Editor*
Cambridge Companions to Philosophy
0-521-65178-6 Hardback $54.95
0-521-65695-8 Paperback $19.95

A Semiotic Theory of Theology and Philosophy
Robert S. Corrington
0-521-78271-6 Hardback $59.95

Appropriating Heidegger
James Faulconer *and*
Mark Wrathall, *Editors*
0-521-78181-7 Hardback $59.95

Science and Selection
Essays on Biological Evolution and the Philosophy of Science
David L. Hull
Cambridge Studies in Philosophy and Biology
0-521-64339-2 Hardback $54.95
0-521-64405-4 Paperback $19.95

Aristotle's Philosophy of Biology
Studies in the Origins of Life Science
James G. Lennox
Cambridge Studies in Philosophy and Biology
0-521-65027-5 Hardback $64.95
0-521-65976-0 Paperback $22.95

The Critique of Theological Reason
James P. Mackey
0-521-77293-1 Hardback $59.95

The Cambridge Companion to Malebranche
Steven Nadler, *Editor*
Cambridge Companions to Philosophy
0-521-62212-3 Hardback $54.95
0-521-62729-X Paperback $19.95

The Fragility of Goodness
Luck and Ethics in Greek Tragedy and Philosophy
Updated Edition
Martha C. Nussbaum
0-521-79126-X Hardback $90.00
0-521-79472-2 Paperback $27.95

Women and Human Development
The Capabilities Approach
Martha C. Nussbaum
0-521-66086-6 Hardback $24.95

Justification and Legitimacy
Essays on Rights and Obligations
A. John Simmons
0-521-79016-6 Hardback $54.95
0-521-79365-3 Paperback $19.95

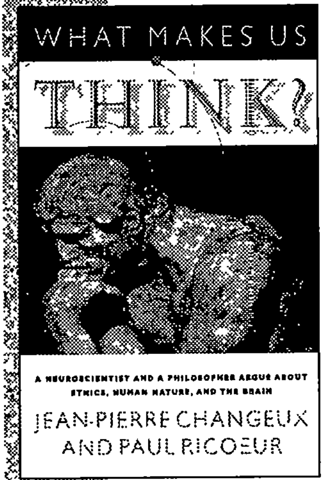
Thomas Reid and the Story of Epistemology
Nicholas Wolterstorff
Modern European Philosophy
0-521-79013-1 Hardback $54.95

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

40 West 20th Street, N.Y., NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cambridge.org
AmEx/MasterCard/VISA accepted. Prices subject to change.

New from Princeton


WHAT MAKES US
THINK?
A NEUROSCIENTIST AND A PHILOSOPHER ARGUE ABOUT ETHICS, HUMAN NATURE, AND THE BRAIN
JEAN-PIERRE CHANGEUX
AND PAUL RICOEUR

What Makes Us Think?

A Neuroscientist and a Philosopher Argue about Ethics, Human Nature, and the Brain

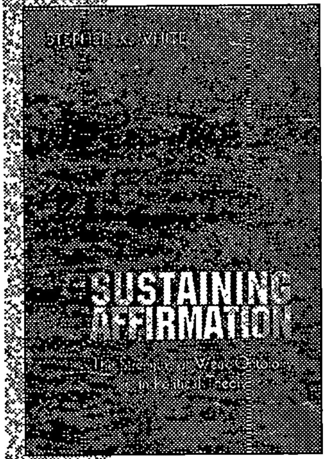
Jean-Pierre Changeux and Paul Ricoeur

Translated by M. B. DeBevoise

Will understanding our brains help us to know our minds? Or is there an unbridgeable distance between the work of neuroscience and the workings of human consciousness? In a remarkable exchange between neuroscientist Jean-Pierre Changeux and philosopher Paul Ricoeur, this book explores these questions.

"These two amazing minds at work make for a fascinating look at the who, what, and how thought happens."
—*Choice*

Cloth $29.95 ISBN 0-691-00940-6



Sustaining Affirmation

The Strengths of Weak Ontology in Political Theory

Stephen K. White

In light of many recent critiques of Western modernity and its conceptual foundations, the problem of adequately justifying our most basic moral and political values looms large. Without recourse to traditional ontological or metaphysical foundations, how can one affirm—or sustain—a commitment to fundamentals? The answer, according to Stephen White, lies in a turn to "weak" ontology.

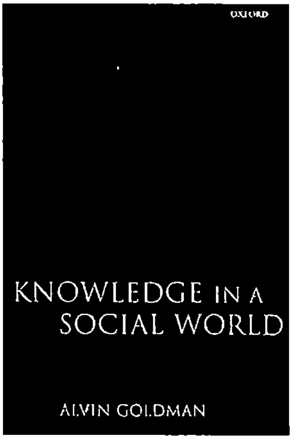
Paper $15.95 ISBN 0-691-05033-3
Cloth $49.50 ISBN 0-691-05032-5

Princeton University Press
800-777-4726 • WWW.PUP.PRINCETON.EDU

do not print this information J. OF PHILOSOPHY 11/00 Q09 01-139

KNOWLEDGE IN A SOCIAL WORLD

ALVIN I. GOLDMAN



"An important work of monumental scope. Its central concern is the ways in which social practices and familiar institutions contribute to and undermine the pursuit of knowledge.... A book of singular importance to lawyers, political theorists, social and natural scientists as well as to educators and theorists of education. It is a major contribution to all these fields and not just because of its insights into them, but for its accessibility to intelligent practitioners as well. It may well be the most significant interdisciplinary philosophy book of the decade."

—Jules Coleman, *Yale Law School.*

"Alvin Goldman, the premier epistemologist of the past two decades, has written a pioneering book that will define the field of social epistemology. Scholars will learn from his judicious and lucid proposals, and they will be wrestling for years with the exciting and important problems he raises."

—Philip Kitcher.

"...belongs to a sparse tradition in epistemology given to the epistemic evaluation of social institutions. Landmarks in this tradition include Bacon's *The Advancement of Learning*, Locke's *Some Thoughts Concerning Education*, and Dewey's *Social Inquiry*. Goldman's book is the equal of these influential volumes in breadth of vision, mastery of relevant contemporary thought, and practical utility, and surpasses them in theoretical depth. It is a work crafted with supreme philosophical skill."

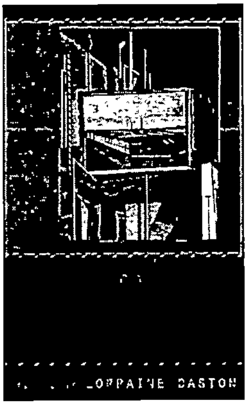
—Frederick Schmitt.

Knowledge in a Social World offers a philosophy for the information age. Alvin Goldman explores new frontiers by creating a thoroughgoing *social* epistemology, moving beyond the traditional focus on solitary knowers. Against the tides of postmodernism and social constructionism Goldman defends the integrity of truth and shows how to promote it by well-designed forms of social interaction. From science to education, from law to democracy, he shows why and how public institutions should seek knowledge-enhancing practices. The result is a bold, timely, and systematic treatment of the philosophical foundations of an information society.

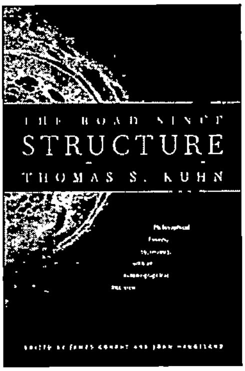
1999 424 pp.; 3 line illus paper $19.95 cloth $70.00

OXFORD UNIVERSITY PRESS

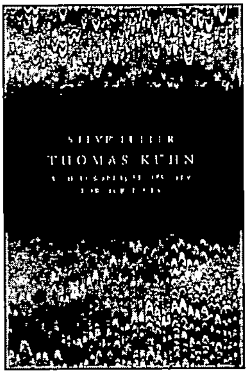
the life of science



Biographies of Scientific Objects
Edited by Lorraine Daston
"Why do scientists select or create some things but not others to be the objects of their investigations? This impressive and highly original collection of essays, ranging from sociology and anthropology to theoretical physics and mathematics, looks at the varied and often surprising processes—intellectual, practical, and social—through which the subject matters of the sciences are brought into being, and, on occasion, may pass away. Biographies of Scientific Objects is a tour de force."—Nicholas Jardine, University of Cambridge
Paper $19.00



The Road since Structure
Philosophical Essays, 1970–1993, with an Autobiographical Interview
Thomas S. Kuhn
Edited by James Conant and John Haugeland
This volume is the fullest record we now have of the new direction Thomas Kuhn was taking during the last two decades of his life. The first part of the book consists of essays in which Kuhn refines the basic concepts set forth in *The Structure of Scientific Revolutions,* and in part two, Kuhn replies to many of the criticisms of his earlier work. The third part is the transcript of a remarkable and singular interview that the normally reticent Kuhn conducted just before his death.
Cloth $25.00



Thomas Kuhn
A Philosophical History for Our Times
Steve Fuller
"The resonance of the phrase 'paradigm shift' amply testifies to the profound influence of Thomas Kuhn's book *The Structure of Scientific Revolutions* (1962). For Fuller, that influence cannot end soon enough. . . . A brilliant analysis deserving as wide a readership as the acclaimed book it critiques."—Booklist, starred review
Cloth $35.00

The University of Chicago Press 5801 South Ellis Avenue, Chicago, IL 60637
www.press.uchicago.edu

OXFORD UNIVERSITY PRESS

KANT'S THEORY OF MIND

An Analysis of the Paralogisms of Pure Reason

Second Edition

KARL AMERIKS

This seminal contribution to Kant studies, originally published in 1982, has now been brought up to date with a substantial preface and postscript.

2000 392 pp. cloth $65.00 paper $19.95

CULTURAL OTHERNESS

Correspondence with Richard Rorty

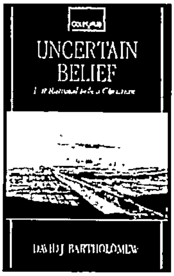
Second Edition

ANINDITA NIYOGI BALSLEV

"The letters...are a fine example of what they advocate—genuine dialogue between real interlocutors, and not fantasies of what some 'other' might say."
—Cleo McNelly Kearns, Rutgers University.

(AAR Cultural Criticism Series 4)
(An American Academy of Religion Book)

2000 128 pp. paper $16.95



New in paperback!

UNCERTAIN BELIEF

Is it Rational to be a Christian?

DAVID J. BARTHOLOMEW

"Recommended for all college collections."—*Choice*. Arguing from a probabilistic perspective, Bartholomew provides an answer to the fundamental question: is it rational to be a Christian?

1996 (paper 2000) 304 pp.
paper $19.95 cloth $35.00

BERNARD OF CLAIRVAUX

G. R. EVANS

In this book the renowned medievalist G.R. Evans provides a concise introduction to St. Bernard of Clairvaux (1090-1153). The only available introduction to Bernard's life and thought, this latest addition to the *Great Medieval Thinkers* series will appeal to a wide audience of students and scholars of history and theology.

(Great Medieval Thinkers)

2000 232 pp. paper $18.95 cloth $45.00

New in paperback!

VIRTUAL REALISM

MICHAEL HEIM

"Heim's views on technology are far removed from the routine rhetoric of cyberhype. With one foot in the book-crammed world of the humanist scholar and the other in the factories of the future, his ideas are always sensitive to both the euphoria and the anxiety which new technologies engender. Above all, his approach is pragmatic: accepting our technological fate and examining different ways to absorb its impact."—Mark Nixon, *Twenty First Century*.

1998 (paper 2000) 264 pp.; 22 4-color illus, 16 b/w illus paper $16.95 cloth $26.00

New in paperback!

THE SCENES OF INQUIRY

On the Reality of Questions in the Sciences

NICHOLAS JARDINE

"Offers many refreshing and sensible insights to current debates about truth, scientific realism, and particularly the historical evaluation of scientific progress."
—*Philosophical Review*.

1991 (paper 2000) 328 pp. paper $24.95

INDETERMINATE IDENTITY

Metaphysics and Semantics

TERENCE PARSONS

Parsons presents a lively and controversial study of philosophical questions about identity. He claims that there is genuine indeterminacy of identity in the world, articulating such a view in detail and defending it from a host of criticisms.

2000 210 pp.; 26 figs. $49.95

The final volume of Ward's acclaimed tetralogy!

RELIGION AND COMMUNITY

KEITH WARD

This book explores the relationship between religion and society, and discusses the ways in which the major world religions need to adapt to the modern world. Keith Ward looks at different forms of religious community, then proposes a radical vision of the church as a person-affirming, world-transforming society within the emerging global community.

2000 376 pp. paper $19.95 cloth $78.00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, 1-800-387-8020. **www.oup.com**

new titles in philosophy

The MIT Press

Dennett's Philosophy

A Comprehensive Assessment

edited by Don Ross, Andrew Brook, and David Thompson

Essays that consider Dennett's work on evolution, intentionality, consciousness, ontology, ethics and free will.

A Bradford Book
448 pp. $27.95 paper

The Mind Doesn't Work That Way

The Scope and Limits of Computational Psychology

Jerry Fodor

"This book is as important as *Modularity of Mind* was almost twenty years ago."
— Jacques Mehler, Centre National de Recherche Scientifique, France, and International School for Advanced Studies, Trieste, Italy

A Bradford Book
Representation and Mind series
144 pp., 1 illus. $22.95

On the Pragmatics of Social Interaction

Preliminary Studies in the Theory of Communicative Action

Jürgen Habermas

These five lectures, from a period when Habermas was just organizing his thoughts on communication and society, offer an excellent introduction to his work, focusing on the germs of the ideas he would elaborate in *The Theory of Communicative Action.*

Studies in Contemporary German Social Thought
205 pp. $24.95

The Deleuze Connections

John Rajchman

"Anyone who thinks Deleuze is difficult should read this extraordinary book."
— D. N. Rodowick, Chair in Film Studies, King's College, London, and author of *Gilles Deleuze's Time Machine*

120 pp. $14.95 paper

Consciousness, Color, and Content

Michael Tye

"A splendid book...required reading for anyone interested in philosophical theories of consciousness, mental representation, or perception."
— Alex Byrne, Department of Linguistics and Philosophy, MIT

A Bradford Book
Representation and Mind series
184 pp., 8 illus. $26.95

Bioethics

Ancient Themes in Contemporary Issues

edited by Mark G. Kuczewski and Ronald Polansky

"A must read for those in bioethics."
— Patricia Werhane, Ruffin Professor of Ethics, University of Virginia

A Bradford Book
Basic Bioethics series
310 pp. $34.95

Metaphor in Context

Josef Stern

"A remarkable piece of work, broad in scope, artful, and a great read."
— Stephen Neale, Professor of Philosophy, Rutgers University

Studies in Contemporary German Social Thought
464 pp. $18.95 paper

Coherence in Thought and Action

Paul Thagard

"[W]ill be read by philosophers interested in almost any aspect of the subject, and more generally by anyone interested in cognitive science." — Gilbert Harman, Professor of Philosophy, Princeton University

A Bradford Book
Life and Mind: Philosophical Issues in Biology and Psychology
339 pp., 32 illus. $32.95 paper

now in paperback

Oratio Obliqua, Oratio Recta

An Essay on Metarepresentation

François Recanati

"[O]ffers a distinctive and original perspective on central problems in the philosophy of language."
— Robert Stalnaker, Department of Linguistics and Philosophy, MIT

A Bradford Book
Representation and Mind series
450 pp., 2 illus. $24.95 paper

A Zone Book

Detour and Access

Strategies of Meaning in China and Greece

François Jullien
translated by Sophie Hawkes

An exploration of indirect and allusive modes of expression in ancient Chinese culture.

Distributed for Zone Books
435 pp. $32

Sublime Understanding

Aesthetic Reflection in Kant and Hegel

Kirk Pillow

"[O]ne of the most useful books on Kant's aesthetics to have appeared in a decade, and the most fruitful juxtaposition of Kantian and Hegelian aesthetics to have appeared in a generation."
— Paul Guyer, Murray Professor in the Humanities, University of Pennsylvania

Studies in Contemporary German Social Thought
386 pp. $37.95

Beyond Hegel and Nietzsche

Philosophy, Culture, and Agency

Elliot L. Jurist

This book places Hegel and Nietzsche in conversation with each other, reassessing their relationship in a way that affirms its complexity.

Studies in Contemporary German Social Thought
394pp. $37.95

German Realists in the Nineteenth Century

Georg Lukács
translated by Jeremy Gaines and Paul Keast
edited and with an introduction and notes by Rodney Livingstone

Literary studies on major German writers of the 19th century by one of the most controversial Marxist philosophers of the 20th century.

396 pp. $24.95 paper

Endgames

The Irreconcilable Nature of Modernity
Essays and Lectures

Albrecht Wellmer
translated by David Midgley

Includes essays on political philosophy, the continuing legacy of the Frankfurt school, architecture, and a number of Wellmer's dialogue partners, from Adorno to Arendt.

Studies in Contemporary German Social Thought
368 pp. $24.95 paper

Reality Transformed

Film as Meaning and Technique

Irving Singer

"Taking as his premise the currently unfashionable idea that narrative cinema has something to do with life, Singer demonstrates how it transforms our common experience into art."
— *Sight & Sound*

232 pp., 3 illus. $16.95 paper

A Zone Book

A Thousand Years of Nonlinear History

Manuel De Landa

"Forcefully challenges habituated understandings of 'history', 'urban' and 'economics'." — Christopher Hight, *AA Files*

Distributed for Zone Books
333 pp., 10 illus. $16 paper

To order call **800-356-0343** (US & Canada) or 617-625-8569. Prices subject to change without notice.
http://mitpress.mit.edu

New from MINNESOTA

TOM COHEN, BARBARA COHEN,
J. HILLIS MILLER, AND
ANDRZEJ WARMINSKI, EDITORS

Material Events

Paul de Man and the Afterlife of Theory

Responding to the problematic of "materialism" as posed in Paul de Man's posthumous last book, *Aesthetic Ideology*, a diverse and distinguished group of scholars explores the question of "material events" in ways that illuminate not just de Man's work but their own.

$22.95 Paper $57.95 Cloth 416 pages
Available December

ISABELLE STENGERS

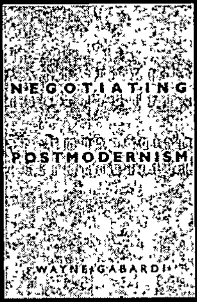
The Invention of Modern Science

Isabelle Stengers, one of the most important and influential philosophers of science in Europe, suggests that we might understand the tension between scientific objectivity and belief as a necessary part of science, central to the practices invented and reinvented by scientists.

$19.95 Paper $49.95 Cloth 184 pages
Available November
Theory Out of Bounds Series, volume 19

WAYNE GABARDI

Negotiating Postmodernism



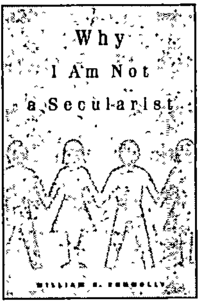
Joining the modern-postmodern debate as it arrives at a critical juncture, this book suggests that the polarizing polemics of the radical postmodernists who once dominated the discussion have given way to a new "critical postmodernism" characterized by dialogue, accommodation, and synthesis.

$16.95 Paper $42.95 Cloth 208 pages
Available December

Now in Paperback!

WILLIAM E. CONNOLLY

Why I Am Not a Secularist



In *Why I Am Not a Secularist*, distinguished political theorist William E. Connolly argues that the secularist's critique of religion's influence in American politics, although admirable in its pursuit of freedom and diversity, too often undercuts these goals through its narrow and intolerant understandings of public reason.

$18.95 Paper 224 pages

University of Minnesota Press
www.upress.umn.edu 773-568-1550

MIN
NE
SO
TA

THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR CAMPUS

VOLUME XCVII, NUMBER 12
DECEMBER 2000

page
635 *Multiple Realizations* Lawrence A. Shapiro
655 *On Explaining That* Paul M. Pietroski

COMMENTS AND CRITICISM

663 *Measuring Confirmation and Evidence* Ellery Eells and Branden Fitelson

BOOK REVIEWS

673 Ian Hacking: *The Social Construction of What?* John Dupré
677 Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld: *Rethinking the Foundations of Statistics* Henry E. Kyburg, Jr.

681 INDEX TO VOLUME XCVII

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Achille C. Varzi. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

2000

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2000 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVII, NO. 12, DECEMBER 2000

MULTIPLE REALIZATIONS*

> We could be made of Swiss cheese and it wouldn't matter.[1]

> For if one defines the operation of sawing as being a certain kind of dividing, then this cannot come about unless the saw has teeth of a certain kind; and these cannot be unless it is of iron.[2]

Philosophers are near consensus that the multiple realizability of higher-level properties in lower-level properties stops theoretical reduction dead in its tracks. Hilary Putnam introduced the idea of multiple realizability in a series of papers in the 1960s,[3] and used it to demolish the hypothesis that psychological states are reducible to brain states. In the following decade, Jerry Fodor[4] generalized Putnam's conclusion and played the *multiple realizability thesis* (henceforth MRT) like a trump card against attempts to reduce *any* special science, that is, any science "above" physics. Whereas a few philosophers have questioned whether MRT is really at odds with reductionism—Robert Richardson[5] and Elliott Sober[6] argue that

---

*For valuable comments and discussion, I wish to thank Tom Bontly, Berent Enç, Branden Fitelson, Mohan Matthen, Jodi Supanich, and, especially, Elliott Sober.

[1] Hilary Putnam, *Mind, Language, and Reality: Philosophical Papers,* Volume II (New York: Cambridge, 1975), p. 291.

[2] Aristotle, *Physics*: Book II, in Richard McKeon, ed., *The Basic Works of Aristotle* (New York: Random, 1941), chapter 9, p. 252.

[3] For instance, in 1967 Putnam published both "The Mental Life of Some Machines" and "The Nature of Mental States." Both are reprinted in his *Mind, Language, and Reality,* pp. 408-28, 429-40.

[4] "Special Sciences," as reprinted in his *Representations* (Cambridge: MIT, 1981), pp. 127-45.

[5] "Functionalism and Reductionism," *Philosophy of Science,* XLVI, 4 (December 1979): 533-58.

[6] "The Multiple Realizability Argument against Reductionism," *Philosophy of Science,* LXVI, 4 (December 1999): 542-64.

0022-362X/00/9712/635–54

MRT is consistent with reductionism—no one, to my knowledge, has seriously questioned the truth of MRT itself. Presumably, this complacency speaks to the obvious nature of MRT. Surely, there are many ways to build a system with a given function, whether the system falls within the domain of psychology, biology, or chemistry. Dissenters focus on what this fact means for reductionism.

Despite philosophers' ready acceptance of MRT, it is not a precise thesis. Indeed, as far as I know, no philosopher has ever tried to complete the sentence, '*N* and *M* are distinct realizations of *T* when and only when _________'. Rather, philosophical discussion of MRT is dominated by examples that appeal to the intuitive force of the thesis. We are asked to imagine kinds, like *carburetors, mousetraps, computers,* and *minds,* that all, presumably, lend themselves to multiple realizations. But why should we accept so easily that MRT is as plausible for minds as it is for carburetors? Furthermore, how much similarity must realizers bear to each other to justify the claim that they are realizations of a single kind? Are corkscrews that differ in composition only—one is made of steel and the other of aluminum—different realizations of the same kind? Are corkscrews which are identical in composition but which differ only in color—one is painted yellow and the other red—different realizations of a corkscrew? Are corkscrews that differ in mechanism as well as in composition—a waiter's corkscrew relies on a simple lever whereas a "winged" corkscrew utilizes a rack and pinion—different realizations of the same kind? What do we say about devices that open wine bottles with injections of CO_2? Perhaps these are not realizations of corkscrews, because they have no screw, but they and corkscrews are realizations of the kind *bottle opener.* But, of course, neither CO_2 injectors nor corkscrews can open soda bottles, so perhaps we should say that corkscrews and CO_2 injectors are realizations of the kind *corked-bottle opener.* But is this a legitimate kind? Who gets to say?

My bet is that philosophers will disagree about how to answer the questions above; or, if they agree, the reasons for their answers will differ. This fact is significant given the weight MRT bears in arguments against reductionism. Before it is possible to evaluate the force of MRT in arguments against reductionism, we must be in a position to say with assurance what the satisfaction conditions for MRT actually are. It is my contention here that philosophers are much too quick to claim that a given kind is multiply realizable. Once various conceptual issues are clarified, the task of demonstrating multiple realizability smacks the hard surface of empirical fact, and, I shall argue, leaves MRT far more difficult to establish than philosophers currently acknowledge. Furthermore, genuine cases of multiple realizability undercut the traditional motivation for admitting functional kinds into the ontologies of the special sciences. In the end, reflection on MRT obliges us to articulate anew some of the basic goals of the special sciences.

I. AN A PRIORI ARGUMENT FOR MULTIPLE REALIZABILITY

A good place to begin an evaluation of MRT is with an examination of arguments that philosophers have made in its favor. Why think that it is possible for a kind to be multiply realizable? One argument that seems to lend support for MRT derives from the concept of functional isomorphism. As Putnam[7] defines this concept, "Two systems are functionally isomorphic if there is a correspondence between the states of one and the states of the other that preserves functional relations" (*ibid.*, p. 291). Putnam's definition of functional isomorphism has clearest application in the context of computing machines, where we can speak of the states of one machine being mapped one-to-one onto the states of a second machine. The two machines will exhibit a functional isomorphism if and only if the sequential relations among states in the first mirror the sequential relations among states in the second onto which they are mapped: "for F to be a functional isomorphism, it must be the case that state A is followed by state B in system 1 if and only if state F(A) is followed by state F(B) in system 2" (*ibid.*, p. 292). Given this notion of functional isomorphism, Putnam claims, it follows *immediately* that "two systems can have quite different constitutions and be functionally isomorphic. For example, a computer made of electrical components can be isomorphic to one made of cogs and wheels" (*ibid.*, p. 292). Functional isomorphism, Putnam seems to suggest, implies MRT.

Before considering the strength of this argument, it is worth noting a curious feature that sets it apart from the others we shall examine: it is an a priori argument. This is curious because MRT is a thesis about the physical world. According to MRT, it is physically possible for some undefined class of kinds to have multiple realizations. Thus, one might think, the truth of MRT rests on facts about the way the world is. But then one must judge the truth of MRT on the basis of empirical evidence. The above argument for MRT does no such thing, however. Rather, it appears to proceed from a definition of functional isomorphism to a conclusion about the physical possibility of multiple realizability.

---

[7] "Philosophy and Our Mental Life," in his *Mind, Language, and Reality*, pp. 291-303.

In fact, I think the argument from functional isomorphism does not go *any* distance in establishing the truth of MRT. Then what does the functional isomorphism between computers made with electrical components and computers made of cogs and wheels show? The temptation is to call these two computers multiple realizations of a machine table, that is, a program. But computers are not realizations of machine tables—not in anything like the sense that winged corkscrews and waiter's corkscrews are both realizations of a corkscrew. Although I have not said anything yet about what should count as a realization of a kind, presumably a realization of a kind must exhibit relevant properties of the kind or must, perhaps, have the *function* of exhibiting relevant properties of the kind.[8] For instance, if a mechanism is to count as a corkscrew, it must be able to remove corks or, at least, be intended to remove corks. But computers are not machine tables, nor are they intended to be machine tables. Rather, computers are devices for implementing the sequence of functional relations that a machine table describes. With this point in clear sight, the empirical nature of MRT also returns to view. It is, after all, an empirical question whether a given physical device is capable of preserving the sequence of functional relations that a machine table describes. If we believe that electrical components as well as cogs and wheels are indeed equally suited to maintaining the sequence of functional relations that a given machine table specifies, then we believe this because of our familiarity with the causal properties of electrical components and cogs and wheels.

Let me be clear. I do not deny Putnam's claim that "two systems can have quite different constitutions and be functionally isomorphic."[9] I do, however, deny that two systems that are functionally isomorphic are, in virtue of this fact, realizations of the same kind. In some cases, two constitutionally distinct systems may implement the same program. In this case, the systems would be functionally isomorphic. This does not force us to the conclusion, however, that the systems are realizations of the same kind. Perhaps the program is a very simple one and one of the devices in which it is implemented—a mainframe computer—is capable of running much more sophisticated programs. Suppose now that the other device is a pocket calculator and, due to its very small RAM, is capable of computing only basic arithmetic functions on numbers with three or fewer digits. This seems to me evidence that the two devices, although for the moment functionally isomorphic, are not realizations of the same kind. Furthermore, one could imagine a functional isomorphism between a functioning mousetrap and marks on a slip of paper. *Mouse present* is represented on paper with the symbol '*A*'; *mouse in trap* is represented with the symbol '*B*'; *mouse dead* is represented with the symbol '*C*'; and *mouse removed* is represented with the symbol '*D*'. If one wanted to, one could preserve an isomorphism between the states of the mousetrap and the marks on the slip of paper. I hope everyone would agree, however, that maintenance of such an isomorphism does not, as if by some occult force, transform the paper and printed characters into a mousetrap. The mousetrap and the piece of paper are not both realizations of a single kind (that is, of any *interesting* single kind).

---

[8] Here, I am assuming a difference between machine functionalism and teleological functionalism. Machine functional kinds are defined purely formally, by reference to relations they maintain between inputs, other states, and outputs—see my "Behavior, ISO Functionalism, and Psychology," *Studies in History and Philosophy of Science*, xxv, 2 (1994): 191-209. Teleological functionalists individuate kinds on the basis of their purpose. Granting this distinction, my claim here is that machine-functionally identical systems need not be realizations of the same kind—this is so when they are not teleologically-functionally identical.

[9] "Philosophy and Our Mental Life," p. 292.

How could one suppose that functional isomorphism implies anything about the truth of MRT? A likely diagnosis is that, as computer science has shown, it is possible to describe the operation of a system at a level of abstraction that ignores details about the constitution of the system. Cognitive science, of course, has taken this message to heart and thus has the goal of developing information-processing descriptions of cognitive capacities that abstract from any mention of neurological underpinnings. But the fact that we can describe the operation of a system at a high level of abstraction does not imply that the system is multiply realizable. Again, this depends purely on whether there are physically possible but distinct means by which to produce the properties that define the system. Slips of paper might work as well as ivory for purposes of playing chess, but slips of paper do not catch and kill mice. They cannot. It is simply a confusion to identify a description of sequences of relations among functional states of a system with the system itself.[10] Once we see the distinction between a description of a system and the system described, the temptation to move from claims of functional isomorphism to the truth of MRT loses its allure.

---

[10] See also John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, III (1980): 417-57; and Elliott Sober, "Learning from Functionalism: Prospects for Strong Artificial Life," in Margaret A. Boden, ed., *The Philosophy of Artificial Life* (New York: Oxford, 1996), pp. 361-78.

II. A LIKELIHOOD ARGUMENT FOR MULTIPLE REALIZABILITY

Putnam[11] presents an argument for MRT which is perhaps best reconstructed as a *likelihood argument*. Likelihood arguments involve the comparison of two or more competing hypotheses. The more likely hypothesis is the one that would make more probable the evidence at hand. So, for instance, a poker player in the movie *Rounders*, intent on convincing his girlfriend that success in poker is not merely a matter of luck, draws her attention to the fact that the same faces are present year after year in the World Series of Poker tournament. More explicitly, the poker player is asking his girlfriend to consider two hypotheses:

(L): Success in poker is purely a matter of luck.
(S): Success in poker requires skill.

The question the poker player then puts to his girlfriend is this: Which hypothesis makes the evidence—the fact that the same players make it to the World Series of Poker year after year—more probable? Any reasonable person ought to agree that the evidence is more probable given (S) than (L).

In like fashion, one could read Putnam as asking us to consider two hypotheses:

(TI): Psychological states are (type) identical to physical-chemical states of the brain.
(MR): Psychological states are multiply realizable.

Putnam then asks us to consider which hypothesis would make more probable that mammals, reptiles, mollusks, and extraterrestrials all share a single type of psychological state, for example, pain. If (TI) were true, we should be very surprised to find that mammals, reptiles, and the like share a single psychological state because this would imply that they share a single kind of physical-chemical state. But, from what we know about the brains of mammals, reptiles, and mollusks, and from what we *do not* know about the brains of extraterrestrials, we have ample reason to be skeptical that all forms of life that feel pain also share a single kind of physical-chemical state. On the other hand, acceptance of (MR) makes identity of psychological states across life forms, if not probable, at least not as surprising as (TI) makes it. (MR) is thus the likelier of the two hypotheses.

The likelihood argument for MRT is compelling, although perhaps not so much as the poker player's argument. In particular, whereas we know something about the odds of holding particular hands in card games, we know very little about the properties of the brain which make it a realization of a mind. Perhaps there *is* only one physically possible physical-chemical state that can realize pain. What is surprising about this? Consider instead the property of *being able to scratch a sapphire.* On Mohs's scale of hardness, a sapphire ranks just below a diamond, which is the hardest substance on the scale. Suppose it is true that diamonds are the only physically possible substance capable of scratching sapphires. Granting this, then diamonds are the only possible realizer of the property of being able to scratch a sapphire. Because MRT is an empirical thesis, we must be willing to accept on occasion that it might, as a matter of empirical fact, be false. Hence, even though hypothesis

---

[11] "The Nature of Mental States."

(MR_S): Scratchers of sapphires are multiply realizable.

may seem likelier than

(D): Scratchers of sapphires are (type) identical to diamonds.

(MR_S) is, we might suppose, false. But then we must ask why we should accept (MR) as more likely than (TI).

Indeed, Putnam acknowledges that

> It is not altogether impossible that such a [physical-chemical] state will be found. Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal.... Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical 'correlate' of pain (*ibid.*, p. 436).

Perhaps sensibly, Putnam is dubious of this suggestion, claiming that it is "certainly an ambitious hypothesis" (*ibid.*). Nevertheless, it is a point worth bearing in mind for the simple reason that it returns our attention to the empirical nature of MRT. In particular, it is a point that invites caution when citing MRT as an obstacle to reduction. Maybe Fodor's (*op. cit.*) classic defense of the special sciences is not as general as he would have us believe. It is quite possible that some special sciences are more susceptible to reduction than others, depending on how the world turns out to be. I shall have more to say on the issue of reduction below.

III. EMPIRICAL EVIDENCE FOR MULTIPLE REALIZABILITY

Ned Block and Fodor[12] agree with Putnam that the "argument against physicalism rests upon the empirical likelihood that creatures of different composition and structure, which are in no interesting sense in identical physical states, can nevertheless be in identical psychological states" (*ibid.*, p. 80). Rather than present a likelihood argument, however, Block and Fodor appeal to "three kinds of empirical considerations" (*ibid.*) which, they think, provide direct evidence of multiple realizability. Amusingly, one source of evidence they cite is convergent evolution—the same phenomenon that Putnam takes to be the best reason to *doubt* MRT. Block and Fodor think that physiologically distinct organisms might have hit upon identical psychological solutions to various environmental problems, and thus exhibit multiple realizations of identical psychological capacities. Ought one, as Putnam does, take the mammalian eye and the mollusk eye as "virtually identical" realizations of an eye, or, as Block and Fodor seem inclined to suggest, treat them as distinct realizations of an eye? I began here by claiming that philosophers have been too serene in their acceptance of MRT and here we see a case in point. Disputes like these must be recognized, diagnosed, and resolved before we are in a position to decide what threat if any MRT brings to reduction. Below, I take steps toward this end.

---

[12] "What Psychological States Are Not," as reprinted in Fodor, *Representations* (Cambridge: MIT, 1981), pp. 79-99.

Block and Fodor cite two other lines of evidence in favor of MRT. The first is the "Lashleyan doctrine of neurological equipotentiality [which] holds that any of a wide variety of psychological functions can be served by any of a wide variety of brain structures" (*ibid.*). More recently, the Lashleyan doctrine has given way to the more modest idea that the brain is surprisingly plastic. Simply put, it appears that the same psychological state may be token identical to different kinds of brain states in different people, or different kinds of brain states at different times in a single person. If this is true, it seems a strong argument in favor of MRT. Secondly, Block and Fodor ask us to grant the "conceptual possibility that psychological predicates could apply to artifacts" (*ibid.*, p. 81). Because, they believe, it is "very likely" that we may someday build machines with minds, we must accept MRT. After all, machines, by hypothesis, are physiologically distinct from us, and thus the possibility of machine minds suffices to establish MRT.

Let us suppose that everything Block and Fodor say is true. Psychologically identical states can be identified with different parts of the brain in different or even the same individuals. Natural selection has figured out different physiological means to the same psychological ends. Artificial intelligence is a conceptual possibility. Indeed, let us throw into this collection of evidence the conceptual possibility of extraterrestrial intelligence. The question we must now ask is whether this establishes MRT.

IV. EVALUATING THE EMPIRICAL EVIDENCE FOR MULTIPLE REALIZABILITY

As is now apparent, philosophers tend to associate MRT with the issue of mind-brain reduction. As Fodor (*op. cit.*) argues, however, MRT is a thesis that is supposed to have consequences for any reductionist program. It may be that we are in a better position to evaluate the force of MRT for some reductionist efforts than we are for others. I shall now discuss some conceptual issues that advocates of MRT must address prior to considering the implications of MRT for particular reductions. These issues, because they are general, naturally have consequences for the topic of mind-brain reduction as well. I want to emphasize, however, that my goal here is not to defend a theoretical reduction between psychology and neurobiology. My present aim is to reveal some hitherto unappreciated difficulties in evaluating evidence often thought to establish MRT.

Let us begin with a distinction that emerges from reflection on the kinds of things for which MRT has a chance of being true—the kinds of things that may be multiply realizable. MRT, to the extent that it is true, is true of kinds that are defined by reference to their purpose or capacity or contribution to some end.[13] Purposes, capacities, contributions—these specify roles to be played and, accordingly, it may be possible that any of a variety of occupants can fill the role. Philosophical discussion of multiple realizability has made frequent reference to carburetors, mousetraps, computers, and minds because these are all kinds that we identify in virtue of what they do. In contrast, water is a kind defined not by what it does but by its molecular structure. Whatever is H_2O is water, and whatever is not is not. 'Water' is not a term that implies a commitment to any particular activity or disposition or goal, and so water is not a candidate for multiple realizability. Of course, if we choose to characterize water in terms of its dispositions, if 'water' is just short for a description—that which freezes at 0 degrees Celsius, boils at 100 degrees Celsius, has a specific gravity of 0, and so on—then we are free to take water to be multiply realizable.

We now come to a distinction that is significant for assessing the strength of empirical evidence for MRT. Some of the properties of the realizers of multiply realizable kinds are relevant to the purpose, activity, or capacity that define the kind and some are not. I asked in the introduction whether two corkscrews, alike in constitution and mechanism but distinct in color, count as alternative realizations of the kind *corkscrew.* I claim that they do not because the only property by which they differ—color—is not a property that contributes to their capacity to remove corks. To say that a kind is multiply realizable is to say that there are *different* ways to bring about the function that defines the kind. But, if two particulars differ only in properties that do not in any way affect the achievement of the defining capacity of a kind, then there is no reason to say that they are tokens of different realizations of the kind. Differently colored corkscrews, alike in every other respect, are not tokens of different realizations of a corkscrew because differences in color make no difference to their performance as a corkscrew. The moral of this example is that multiple realizations count truly as *multiple* realizations when they differ in causally relevant properties—in properties that make a difference to how they contribute to the capacity under investigation.

---

[13] See, for example, William G. Lycan, "Form, Function, and Feel," this JOURNAL, LXXVIII, 1 (January 1981): 24-50.

This moral makes sense of why the waiter's corkscrew and the winged corkscrew do seem to count as multiple realizations of a corkscrew. The waiter's corkscrew relies on a lever to pry the cork out of the bottle whereas the winged corkscrew uses a rack and two pinions to do the same job. Levers and rack and pinions are different mechanisms that require different manipulations, they are described by different laws, and so on. The causally relevant properties of these two devices differ; a fortiori they qualify as different realizations of a corkscrew.

What are we to say about two waiter's corkscrews—one of which is composed of steel and the other of aluminum? On the one hand, steel and aluminum are different materials. If differences in constitution suffice to mark differences in kinds of realizations, then we must say that the steel and aluminum waiter's corkscrews are tokens of distinct kinds of realizations. This conclusion would be too hasty, however. We must first ask what it is about steel and aluminum which contributes to cork removal. Do they differ in their causally relevant properties? They do not. Indeed, it is because they have the *same* causally relevant properties (for example, rigidity) that each serves so effectively in its role as a waiter's corkscrew. Steel and aluminum are *not* different realizations of a waiter's corkscrew because, relative to the properties that make them suitable for removing corks, they are identical. The fact that one corkscrew is steel and the other aluminum is no more a reason to characterize them as different realizations than the fact that one might be yellow and the other red.

Of course, this point does not imply that steel and aluminum *never* qualify as alternative realizations of a kind. Relative to some kinds they may be. For all I know, rigidity is a disposition that is multiply realizable. If rigidity is a disposition that can be brought about in various ways, and if steel and aluminum differ in respect to how they produce rigidity, then steel and aluminum are alternative realizations of rigidity. But, relative to their function in a corkscrew, rigidity *screens off* the differences between steel and aluminum. All that counts in the appraisal of their differences qua corkscrew is how they contribute to cork removal, and here there *is* no difference. The means by which steel and aluminum achieve rigidity are, in this context, beside the point.

The general lesson is this. Showing that a kind is multiply realizable, or that two realizations of a kind are in fact distinct, requires some work. In particular, one must show that the realizations in question differ in causally relevant ways. Earlier, I quoted Block and Fodor as saying that mind-brain reduction is unlikely because almost certainly organisms "which are in no interesting sense in identical physical states, can nevertheless be in identical psychological states" (*op. cit.*, p. 80). Bearing in mind the argument above, one ought to begin to wonder how likely this truly is. How interesting must physical similarities be before they constitute a challenge to MRT? A common thought experiment in philosophy of mind asks us to imagine a brain in which neurons are replaced, one by one, with silicon chips. This example is intended to tug our intuitions toward accepting that the mind is multiply realizable. Standard neural brains have minds. If we replace each neuron with a silicon chip, all the while maintaining the organization of the brain—the connections that previously existed between neurons—we ought to end up with a silicon brain that functions just as the ordinary brain did. But why should we suppose that the switch from neurons to silicon chips marks an "interesting" physical difference? If each neuron's contribution to psychological capacities is solely its transmission of an electrical signal, and if silicon chips contribute to psychological capacities in precisely the same way, then the silicon brain and the neural brain are not distinct realizations of a mind. The difference between neurons and silicon chips, in such a case, is no more interesting relative to their contribution to psychological function than would be the difference between the contributions of neurons gray in color and neurons stained purple. Indeed, there seems no more reason to count as distinct realizations of a brain one in which each neuron has been replaced by a silicon chip and one in which each neuron has been replaced by another neuron.

We are now in a position to resolve the disagreement Putnam and Block and Fodor might have about whether the octopus's eye and the mammalian eye count as distinct realizations of an eye. Putnam assumes these eyes are of a kind because each eye is structurally very similar. Each eye has a single lens that causes an inverted image to fall on a retina, where light is then transformed into electrical signals. Because it is the lens and the retina that contribute to the ability of the eye to see, Putnam might shake off as irrelevant the information that the lens of the octopus's eye is composed of one kind of protein and the lens of the mammalian eye of another, or that the two eyes utilize different kinds of visual pigments. The processes that result in the formation and analysis of an image screen off these facts about microcomposition. Block and Fodor, were they to deny that octopi and mammals have the same kind of eye, would apparently not regard these processes of image formation as physically interesting. For Block and Fodor, eyes would count as distinct whenever they differ in molecular composition. But if the line of argument I have pursued above is correct, Block and Fodor would for similar reasons be committed to distinguishing between kinds of corkscrews simply on the basis of their color.

I have argued that the evidence philosophers casually cite in favor of MRT must meet more stringent standards than philosophers have realized. More specifically, I have argued that differences in physical composition need not imply differences in realization. To establish MRT, one must show that the differences among purported realizations are causally relevant differences. This, I suppose, is an empirical enterprise and its difficulty no doubt varies considerably across subject matters. It may be a trivial matter to devise corkscrews that operate according to distinct physical principles, but how does one show that the physical principles that determine minds can be diverse? Without considerable knowledge about how the brain produces a mind, why be so sanguine that there are in fact many *truly* distinct ways to build a mind? In the concluding section, I shall draw out some further consequences of these ruminations for the issue of theoretical reductionism. Now, however, it is time to examine another question that advocates of MRT must consider.

V. A DILEMMA FOR MULTIPLE REALIZABILITY

I argued above that two realizations of a kind *T* are in fact different kinds of realizations of *T* only when they differ in their causally relevant properties, that is, the properties by which they contribute to the capacity, purpose, goal, and the like that serves to individuate *T* as the kind that it is. On the basis of this principle, I dismissed the claim that two corkscrews that differ only in color are in fact different kinds of realizations of a corkscrew. Similarly, I rejected the suggestion that two waiter's corkscrews that differ only in material composition are distinct realizations of a corkscrew. I allowed, however, that waiter's corkscrews and winged corkscrews are legitimate instances of multiple realizations of a corkscrew. Likewise, I agree with Putnam that the octopus eye and the mammalian eye are not different realizations of an eye. Perhaps, we should say, the camera eye that we share with octopi and the compound eye present in many arthropods are truly different realizations of the kind *eye*. The justification for this claim is that camera eyes and compound eyes "do the same thing" but in very different ways. The causally relevant properties that camera eyes and compound eyes exhibit—those properties in virtue of which they see—are quite distinct and so these two kinds of eyes are in fact *two* kinds of eyes.

But a hard question now looms. Granting what I have said about the considerations that one must bring to bear on evidence for multiple realizability, why should we think that MRT is a coherent thesis at all? The problem is this. Take what appears to be a legitimate case of multiple realization: the waiter's and winged corkscrews or the camera and compound eyes. Either the realizing kinds truly differ in their causally relevant properties, or they do not. If they do not, then we do not have a legitimate case of multiple realizability, and MRT, in this given instance, is false. If the realizing kinds do genuinely differ in their causally relevant properties, then, it seems, they are different kinds. But if they are different kinds, then they are not the same kind, and so we do not have a case in which a single kind has multiple realizations.

Why do we call both the camera eye and the compound eye *eyes*?[14] They are very different kinds of things—constructed of different parts, operating under different optical principles, capable of different powers of resolution. Fodor[15] makes a point that might be helpful in this context: "What justifies a taxonomy, what makes a kind 'natural', is the power and generality of the theories that we are enabled to formulate when we taxonomize in that way" (*ibid.*, p. 119). Fodor makes this remark in the course of defending functional taxonomies, that is, taxonomies that abstract from facts about microcomposition. Following this suggestion, it might appear that the warrant for conceiving of camera eyes and compound eyes as multiple realizations of the kind *eye* is the benefit that doing so has for the science of vision. Despite causally relevant differences between realizations of eyes there are, presumably, interesting similarities. A theory of eyes, Fodor's idea seems to be, would capture these similarities and express them in its laws.

---

[14] It is in fact a gross simplification to think of camera eyes and compound eyes as two well-defined kinds of *eye*. There is tremendous diversity within these two kinds of *eye* and there are some eyes that do not fit comfortably into either category. See, for instance M. Land, "Optics of the Eyes of the Animal Kingdom," in J. Cronly-Dillon and R. Gregory, eds., *Evolution of the Eye and Visual System* (New York: Macmillan, 1991), pp. 118-35.

[15] *Psychological Explanation: An Introduction to the Philosophy of Psychology* (New York: Random, 1968).

Fodor's defense of functional taxonomies is familiar and, some might say, essential if special sciences are to retain their sovereignty. The dilemma I raised above, however, challenges this motivation for functional taxonomies. In particular, the functional taxonomist must explain how, if realizations of a kind are in fact different realizations of a kind, there can be laws under which they are unified. Recall that realizations differ only when their causally relevant properties differ. But if two realizations contribute to a capacity in causally distinct ways, then this must mean that there are no laws common to both of them.[16] Fodor seems prepared for such a charge and seeks to defuse it in the following way:

> Given that water can be "reduced" to H_2O, it is possible to say what all samples of water have in common either in the language of viscosity, specific gravity, and so on at the macrolevel, or in chemical language at the microlevel. It is patent that functional analysis need not share this property of reductive analysis. When we identify a certain mousetrap with a certain mechanism, we do not thereby commit ourselves to the possibility of saying in mechanistic terms what all members of the set of mousetraps have in common. Because it is (roughly) a sufficient condition for being a mousetrap that a mechanism be customarily *used* in a certain way, there is nothing in principle that requires that a pair of mousetraps *have* any shared mechanical properties.[17]

Fodor believes that when a kind can be reduced, it is possible to formulate laws about the kind at both a "higher" and a "lower" level.

---

[16] This claim is perhaps too strong. There may be laws about realizers that refer only to their capacities. Thus, it may be a law that natural selection will select organisms that can maintain their body temperature over those which cannot. This law is mute concerning how organisms maintain their body temperature; for an interesting discussion of the explanatory importance of function attributions, see Berent Enç and Fred Adams, "Functions and Goal Directedness," *Philosophy of Science*, LIX, 4 (December 1992): 635-54. Within many special sciences, however, the name of the game is the discovery of laws about the internal workings of various mechanisms (more on this in the following section). Consider, for instance, that psychology without concern for internal processes is nothing more than behaviorism. The dilemma I describe in this section can, for psychology, be stated as follows: psychology, conceived as the science of the internal workings of the mind, can contain no laws; or psychology must be construed purely behavioristically—as concerned only with laws of association between inputs and outputs.

[17] *Psychological Explanation*, pp. 115-16.

In contrast, because (some?) functional kinds can be multiply realized and are thereby irreducible, there are no laws about them at a lower level. Functional kinds are "visible" only at the level in which we see similarities between otherwise disparate realizations, and so it is at this level that we must look for laws ranging over them.

It seems clear, however, that laws about functional kinds must do more than simply state the capacity in virtue of which a functional kind is the kind that it is. That is, a science of mousetraps ought to be able to tell us more about mousetraps than that they are used to catch mice. It is not a law that all mousetraps are used to catch mice—or if it is a law, then it is an analytic rather than an empirical law, and a numbingly dull law at that.[18] Certainly, it is not anything like the law that all samples of water freeze at 0 degrees Celsius. This law about water derives from empirical investigation and can conceivably be challenged and found to be false. The same cannot be said for the law that all mousetraps are used to catch mice. Given that different realizations of mousetraps have different causally relevant properties, however, what more can be said about mousetraps other than that they are all used to catch mice? Similarly, what do camera eyes and compound eyes have in common other than the fact that they have the function to see? If they share many causally relevant properties, then they are not distinct realizations of an eye. If they have no or only few causally relevant properties in common, then there are no or just a few laws that are true of both of them.

Fodor's defense of functional taxonomies goes wrong in its failure to locate the *source* of macrolevel laws. Consider again Fodor's claim that there exist laws about water at both macro- and microlevels. At the macrolevel, we can identify properties like viscosity, specific gravity, and freezing point which all samples of water must share. At the microlevel, the laws about water make reference to the atomic constituents of water. But these microlevel laws determine the macrolevel laws, in the sense that their truth physically guarantees the truth of the macrolevel laws. Macrolevel laws about water are true because certain microlevel laws about water are true. But, more significantly, macrolevel laws about all samples of water are *possible* because the same microlevel laws are true of all samples of water.

---

[18] Not all analytic laws need be numbingly dull. Sober has reminded me that there are interesting laws concerning fitness despite the fact that these laws follow from the definition of fitness as a propensity to survive and reproduce. For further discussion of this point and a related point about the need to assume the causal completeness of special sciences in which such laws are possible, see Sober's "Physicalism from a Probabilistic Point of View," *Philosophical Studies*, xcv (1999): 135-74

There exist laws about the viscosity, specific gravity, freezing point, and so on of *all* samples of water, because all samples of water are composed of H_2O and are thereby determined, according to the microlevel laws that describe the behavior of H_2O, to exhibit similarities in their viscosity, and so on. Contrast this now with so-called functional kinds like *eyes.* We can say nothing more about what camera eyes and compound eyes have in common other than that they have the function to see; this is because they differ in their causally relevant properties. Because, that is, microlevel laws about camera eyes and compound eyes are very different, there is no reason to expect that there will be any interesting macrolevel generalizations common to both.

In section IV, I suggested a criterion by which to judge whether two realizations are different realizations of a single kind. I argued that realizations ought to be distinguished in kind only if they differ according to how they achieve the capacity that serves to individuate the kind they realize. In this section, I traced a surprising consequence of this criterion. Paradoxically, realizations that meet the above criterion are not, after all, realizations of the same kind precisely because there are no interesting laws that unify them. Taking seriously Fodor's suggestion that we allow power and generality of theory to guide our taxonomic practices, must we now give up the idea that functionalist taxonomies have any scientific value?

VI. MULTIPLE REALIZABILITY AND THE SPECIAL SCIENCES

It may appear at this point that my analysis of MRT has somewhere taken a wrong path—one that tilts steeply downhill and comes to an end in the loose gravel at the canyon's edge. It seems as if I must either deny that the special sciences are autonomous, because higher-level kinds have only a single realization and can thus be reduced, or I must deny that there are empirical laws in special sciences because there are no such laws true of all samples of a functional kind. In other words, either special sciences have no ontological independence from lower-level sciences or, worse, they have no empirical laws, which is just to say that they are not empirical sciences at all. In this final section, I shall defend a characterization of the special sciences that attempts to steer between these undesirable results.

Figure 1, adapted from Fodor's "Special Sciences," illustrates the purported import of MRT for reduction.

P and *Q* are kinds of some higher-level science whereas *a-e* and *j-n* are kinds of some lower-level science. Because *P* and *Q* are supposedly multiply realizable in lower-level kinds, the law $P \rightarrow Q$ cannot be reduced to any single lower-level law. Rather, for each law at the

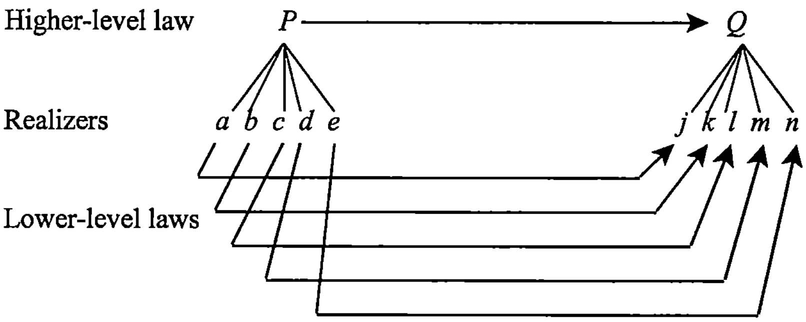


Figure 1

higher level there will be a potentially unlimited number of lower-level laws relating lower-level kinds. My arguments above suggest that Figure 2 or Figure 3 are better representations of the situation.

Figure 2 denies that *a-e* and *j-n* are in fact distinct realizations.

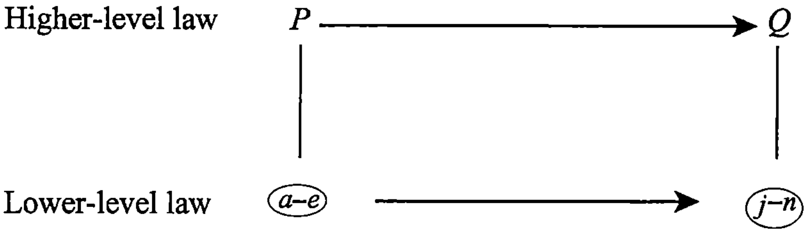


Figure 2

Perhaps *P* is an aquatic camera eye that is capable of producing a focused image. Let *Q* be a spherical lens that possesses a high refractive index at its center which grades off toward its perimeter. This fact that lenses in aquatic camera eyes have graded lenses was discovered by Ludwig Matthiessen in 1877, and it explains how aquatic camera eyes avoid the blurriness that results from spherical aberration.[19] The law $P \rightarrow Q$ then says that all aquatic camera eyes that are capable of producing focused images have graded lenses. If *a-e* are aquatic camera eyes that differ only in physical constitution and *j-n* are graded spherical lenses that also differ only in physical constitution, where none of these differences in physical constitution is causally relevant to the operation of the eye or lens, then, by the argument of section IV, *a-e* and *j-n* are single kinds. The ovals drawn around *a-e* and *j-n* in Figure 2 indicate that, from the perspective of the lower-level science, we are dealing with two kinds of properties rather than two disjunctive sets of properties. Perhaps *a-e* and *j-n* each comprise a single kind within the domain of optics. Law $P \rightarrow Q$ can then be reduced to a law in optics that says something like: all objects that depend on refraction at a curved water-tissue interface for the purpose of focusing an image on a two-dimensional surface require a spherical mechanism the center of which will have a refractive index of roughly 1.52, falling off to a refractive index of 1.4 near the periphery. So, according to Figure 2, it is possible to reduce laws of higher-level sciences (for example, anatomy) to laws of lower-level sciences (for example, optics).

---

[19] Land and Russell Fernald, "The Evolution of Eyes," *Annual Review of Neuroscience*, xv (1992): 1-29.

It is interesting for my purposes to note that at least one preeminent vision researcher seems, at least tacitly, to agree with the picture of reduction I illustrate with Figure 2. Discussing the evolution of graded spherical lenses in aquatic eyes, Russell Fernald[20] notes that "in principle, the identity of the proteins [of which lenses are composed] seems not to be important.... Rather, the distribution of protein concentration as a function of radius is the key to a successful lens" (*ibid.*, p. 257). What matters to whether a spherical lens can focus an image on a two-dimensional surface is that the lens be graded: exactly which kinds of proteins constitute the gradient is beside the point. Suppose the *j-n* in Figure 2 are lenses composed of distinct proteins. Because these lenses, despite their differences in constitution, have identical gradients of refractive indeces, they are, from the perspective of optics, identical lenses. In terms I introduced earlier, a gradient of refractive indeces screens off protein composition from the lens's capacity to focus.

As an alternative to Figure 2, Figure 3 recommends the following redescription of Figure 1. *P* is now a kind that does truly admit of multiple realization.

P, for instance, is the kind *eye* and *a-e* are various eyes that differ in their causally relevant properties. Among *a-e* might be camera eyes, compound eyes, concave reflector eyes that utilize mirrors, scanning eyes that produce a two- or three-dimensional spatial representation from the movement of a one-dimensional detector,[21] and simple

---

[20] "The Evolution of Eyes," *Brain, Behavior and Evolution*, L (1997): 253-59.

[21] Gregory, "Origins of Eyes: With Speculations on Scanning Eyes," in Cronly-Dillon and Gregory, eds., pp. 52-59.

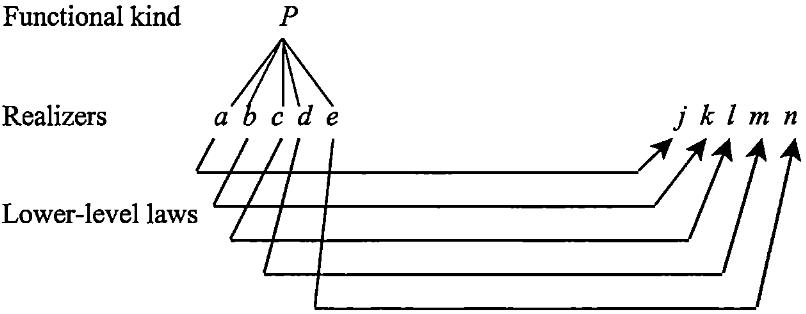


Figure 3

pinhole eyes. By the argument of section v it must be false that there are laws of the form 'All *P*s are also _____'. Rather, at best there will be laws true of each kind that realizes *P*. This, on reflection, seems to be the case. The dramatic diversity among kinds that scientists refer to as *eyes* makes any but analytic statements—for example, all eyes have the function to see—impossible. On the other hand, genuinely empirical laws about various realizers of the kind *eye*—camera eyes, compound eyes, and the like—are common fare in vision science.

What, one may wonder, is the value of categorizing something as a *P* if there are no empirical laws into which *P* enters? It is in response to this question that it becomes necessary to re-examine the goals of the special sciences. In contrast to Fodor's depiction of the special sciences as providing laws about functional kinds, I submit that Figures 2 and 3 provide a more accurate account of the goals of the special sciences. Figure 2 illustrates the special sciences' focus on mechanism. In order to understand why aquatic camera eyes need graded spherical lenses, it is necessary to know something about the optical principles that apply to light and lenses. It is important to realize that this characterization of one aim of the special sciences makes the possibility of reduction an asset rather than a threat. Discovering that an anatomical kind like a camera eye is reducible to particular optical kinds does not make camera eyes disappear. Vision science still has a subject matter. To say that the kinds of vision science can reduce to optical kinds is to say no more than that laws about various kinds that are labeled *eyes* are true in virtue of some deeper set of laws. This bit of news is surely more benefit than bane to the investigator who seeks to understand eyes. Vision scientists can now use their knowledge of optics to explain why particular kinds of eyes have the structures that they do and, indeed, to predict the existence of hitherto undiscovered kinds of eyes.[22] It is difficult to imagine how either of these sorts of tasks would be possible if the laws of special sciences do not reduce to lower-level laws.

Figure 3 captures the special sciences' interest in diversity of mechanism. Despite the fact that we can say nothing more about eyes (in general) than that they have the function to see, it is still a matter of interest what kinds of things can see. We derive a deeper understanding of a given kind of eye from its comparison with other organs that are capable of the same function. Thus, for instance, in trying to understand how a camera eye works it is useful to study its similarities and differences with the compound eye. Doing so provides evidence about which of the properties of the camera eye are causally relevant to the function of seeing and which are not. In short, functional kinds in the special sciences, despite the fact that they never enter into (nonanalytic) laws of the form 'All *P*s are also _____', are of value because they collect and order the domain of a special science in a way that facilitates its investigation.

The points here can be summarized as follows. MRT is not an obvious thesis. The empirical evidence philosophers cite in its favor requires careful scrutiny because kinds that differ in mere constitution may not differ in their causally relevant properties, that is, in the properties by which they contribute to the function that serves to define them. But this standard for multiple realization implies that genuinely different kinds of realizations may not have in common anything *but* their function, and thus there will be no laws true of all realizers qua realizers of a functional kind except laws that are true analytically—for example, all eyes see. Rather than discredit the possibility of special sciences, however, this consequence suggests merely that we abandon a characterization of the special sciences that emphasizes a search for laws over functional kinds. In its place, we should see the special sciences as pursuing an understanding of how functional kinds produce the capacities that make them interesting. Easing this task will be the fact that laws about particular realizations of a functional kind can be reduced. Furthermore, special sciences should continue to use a functional taxonomy of kinds in an effort to group together phenomena that, when laid side-by-side, cast light upon each other, thereby rendering their secrets more visible.

LAWRENCE A. SHAPIRO

University of Wisconsin/Madison

---

[22] Land, *op. cit.*

ON EXPLAINING THAT*

Sentences (1) and (2) below are truth-functionally equivalent, while (2) and (3) are not:

(1) Nora explained why Fido barked.
(2) Nora explained the fact that Fido barked.
(3) Nora explained that Fido barked.

Sentence (3) can be used to report a situation in which Nora uttered 'Fido barked' in response to a question. (Nick asked why the burglar ran off. Nora explained that Fido barked.) In cases of this sort, *that Fido barked* is not the explanandum—the thing explained. Rather, it is what Nora said in the course of explaining something else (for example, the fact that the burglar ran off). But if (1-2) are true, *the fact that Fido barked* is what Nora explained. In this respect, 'the fact that *Q*' is relevantly like 'why *Q*' and not 'that *Q*'. This is somewhat puzzling, especially given the following Fregean view:

> (α) The referent of 'that *Q*' is the proposition expressed by '*Q*'; and facts are true propositions; so if '*Q*' is true, then the referent of 'that *Q*' is the fact that *Q*.

I endorse (α).[1] But even if one rejects (α), one might find it surprising that (2) and (3) have different truth conditions. For how can Nora bear the relation expressed by 'explained' to the fact that Fido barked, without bearing the same relation to the referent of 'that Fido barked'? One cannot, for example, admire the poet Robert Frost without admiring the referent of 'Frost'.

I suggest a semantics for (2-3) that preserves (α) and abandons:

> (β) Instances of '*x* V that *Q*' are true, if and only if *x* bears the relation expressed by (the verb) 'V' to the referent of 'that *Q*'.

If (3) is true, Nora bears *some* relation to the referent of 'that Fido barked', but not a relation expressed by 'explained' (at least not in the ordinary sense of 'expressed'). Drawing on a development of

---

* For helpful comments, thanks to: Mark Baker, Susan Dwyer, Ernie Lepore, and James McGilvray.

[1] For defense, see my *Causing Actions* (New York: Oxford, 2000). Here and throughout, I suppress relativization to context for simplicity.

0022-362X/00/9712/655–62

Donald Davidson's[2] event analysis, I argue that verbs like 'explain' take direct objects of two different sorts, which are associated with two distinct thematic roles. If this is correct, it is at best misleading to speak of the relation expressed by such verbs. This has implications for 'said', 'believes', and so on.

I. PROPOSAL

If (1-3) are true, something *happened.* Nora *acted.* The standard reasons for adopting an event analysis also apply. For example, if x quickly explained why Q, x explained why Q. So one might say the logical forms of (1-3) are: $\exists e$[explain(e, Nora, why Q)]; $\exists e$[explain(e, Nora, the fact that Fido barked)]; and $\exists e$[explain(e, Nora, that Fido barked)]. But this makes it hard to see how (2) and (3) can differ in truth value. In particular, if π is the fact that Fido barked, and π is the referent of 'that Fido barked', then (2) and (3) are both true if and only if $\exists e$[explain(e, Nora, π)]. Suppose, however, that the logical form of (2) is:

(2L) $\exists e$[agent(e, Nora) & explaining(e) & theme(e, the fact that Fido barked)]

where the theme of a Φ-ing is, intuitively and roughly, the thing Φ-ed. By way of comparison, if Nora kicked the ball, $\exists e$[agent(e, Nora) & kicking(e) & theme(e, the ball)]. I assume that certain events have abstract themes. And if x explained the fact that Q, this explanandum is the theme of x's explaining.[3]

According to this kind of event analysis, a verb and its arguments are represented as separate conjuncts of a complex event predicate, and the arguments are associated with *thematic roles.* Davidson originally suggested that an action verb that otherwise appears to have n arguments is true of ordered n+1-tuples, the extra position being for a variable (ranging over events) bound by an existential quantifier; but following Hector-Neri Castañeda,[4] one might suspect that notation like '$\exists e$[kick(e, x, y)]' masks relevant structure. If x kicked y, it *follows* that there was a kicking, and that x did something.

---

[2] "The Logical Form of Action Sentences," in Nicholas Rescher, ed., *The Logic of Decision and Action* (Pittsburgh: University Press, 1967), pp. 87-95. Davidson later endorsed an event analysis, of the sort mentioned in footnote 4 below, that appeals to thematic roles; see his "Adverbs of Action," in Bruce Vermazen and Merrill Hintikka, eds., *Essays on Davidson: Actions and Events* (New York: Oxford, 1985), pp. 230-41.

[3] If x is the agent of an event whose theme is an abstract object y, then perhaps x manipulates some concrete *representation of* y. But I take no stand on this issue here. I allow for events that have multiple themes (and agents), as in 'Three scientists explained six facts'; see Barry Taylor, *Modes of Occurrence* (New York: Oxford, 1985) and Barry Schein, *Plurals* (Cambridge: MIT, 1993). But for simplicity, I assume that the events of explaining in question have unique themes. This is *not,* however, to assume that *facts* have unique explanations. If a single fact can be twice explained, so be it.

More importantly, a range of semantic phenomena—from entailment patterns exhibited by causative verbs to quantificational constructions involving plural noun phases—can be explained by treating action verbs as one-place event sortals, and associating arguments with thematic relations (which hold between events and the semantic values of linguistic expressions that appear in the relevant argument positions).[5]

It may be useful to highlight the contrast between these two eventish proposals in categorical terms. On a traditional view, an intransitive verb like 'yelled' expresses a function of type $\langle i, t\rangle$, from individuals to truth values; a transitive verb like 'kicked' expresses a function of type $\langle i, \langle i, t\rangle\rangle$, from individuals to functions from individuals to truth values. A modification of this view is that 'yelled' expresses a function of type $\langle i, \langle e, t\rangle\rangle$, from individuals to functions from events to truth values; while 'kicked' expresses a function of type $\langle i, \langle i, \langle e, t\rangle\rangle\rangle$, from individuals to functions of the sort expressed by 'yelled'. But on a thematically elaborated event analysis, one would say that 'yelled' and 'kicked' both express functions of type $\langle e, t\rangle$, from events to truth values; although at this point, one might question the value of saying that verbs express functions, as opposed to saying that verbs are satisfied by events. For on the thematically elaborated view, the syntactic arguments of the verb do not denote possible arguments of a function expressed by the verb; 'kicked' is not even true of ordered pairs, much less ordered pairs consisting of a kicker and kickee. From this perspective, intuitions of relationality reflect the fact that 'kicked' is associated with two thematic roles—and thus indirectly associated with ordered pairs $\langle x, y\rangle$ such that x kicked y (since y is the theme of a kicking whose agent is x).

This suggests an account of why (2-3) differ semantically. Suppose the logical form of (3) is:

---

[4] "Comments," in Rescher, pp. 104-12.

[5] For example, see Terence Parsons, *Events in the Semantics of English* (Cambridge: MIT, 1990); Schein, (*op. cit.*), who argues persuasively that certain plural constructions require a thematically elaborated event analysis; Elena Herburger, *What Counts* (Cambridge: MIT, 2000); my *op. cit.* and *Events and Semantic Architecture* (New York: Oxford, forthcoming).

(3L) $\exists e$[agent(e, Nora) & explaining(e) & content(e, that Fido barked)]

where the content of a Φ-ing is, intuitively, the meaning of the Φ-ing. I assume that some explainings have contents, since propositions are expressed in the course of performing the relevant actions. But the content of an explaining is not the explanandum; it is, typically, the explanans. This distinction seems directly relevant to the truth-conditional difference between (2) and (3).

We can also extend this idea to:

(4) The fact that Fido saw the burglar explains the fact that Fido barked.

Suppose that 'explain' has—in addition to its eventive form, which appears in (2-3)—a *stative* form that appears in (4). Using subscripts, let 'explain$_\epsilon$' be a sortal for events, while 'explain$_\sigma$' (like 'knows') is a sortal for states. One can quantify over states as in:

(4L) $\exists s$[subject(s, the fact that Fido saw the burglar) & explaining$_\sigma$(s) & theme(s, the fact that Fido barked)]

On this view, (4) is true if and only if there is a state of explaining whose subject is the fact that Fido saw the burglar, and whose theme is the fact that Fido barked. I assume that states have subjects (not agents), and that to be the subject of a state is to have a certain property. (One might also say that states have objects instead of themes.) But in any case, if (4L) is true, the fact that Fido barked is the explanandum. While the fact that Fido saw the burglar does not act—it does not explain$_\epsilon$ anything—it can explain$_\sigma$ (that is, have the property of explaining$_\sigma$) the fact that Fido barked. The verb 'provide' provides a useful comparison here. If Nora built a well, she may have provided$_\epsilon$ a source of water for someone; but a river can also provide$_\sigma$ a source of water for someone, where 'provide$_\sigma$' is a near synonym to 'be'. Similarly, Nora can provide$_\epsilon$ an answer to a question, while the fact that Fido barked can provide$_\sigma$ an answer.

II. SOME IMPLICATIONS

Setting aside for now these last remarks about states, I have claimed that the logical forms of (2-3) are:

(2L) $\exists e$[agent(e, Nora) & explaining$_\epsilon$(e) & theme(e, the fact that Fido barked)]

(3L) $\exists e$[agent(e, Nora) & explaining$_\epsilon$(e) & content(e, that Fido barked)]

where the theme of an explaining—the explanandum—is distinct from the content of an explaining.

If this is correct, one can grant that (2) and (3) have different truth conditions, while retaining:

> (α) The referent of 'that Q' is the proposition expressed by 'Q'; and facts are true propositions; so if 'Q' is true, then the referent of 'that Q' is the fact that Q.

One can say that the referent of 'that Fido barked' is the proposition expressed by 'Fido barked', and this (true) proposition *is* the fact that Fido barked. On this view, the difference between (2) and (3) lies not in the direct objects having different referents, but in the different thematic relations to a single proposition. A proposition can be the theme *but not the content,* or the content *but not the theme,* of an explaining.

On the other hand, if the proposed treatment of 'explain' is correct, we must reassess:

> (β) Instances of the form 'x V that Q' are true, if and only if x bears the relation expressed by (the verb) 'V' to the referent of 'that Q'.

For (3) is an instance of 'x V that Q'. But it is at least misleading to speak of *the relation expressed* by 'explained$_\epsilon$'. In saying that a word *expresses* a relation, one implies that the word *is true of* (ordered) n-tuples of things. On the thematically elaborated event analysis, however, verbs are true of events or states (not ordered n-tuples). More importantly, there is no *unique* thematic relation associated with 'explained$_\epsilon$'; we can speak of an agent-theme relation or an agent-content relation. And if either relation is primary, it is intuitively the agent-theme relation, which holds between an explainer and the thing explained (the explanandum). But in 'x V that Q', the phrase 'that Q' specifies a content.

One can preserve (β) by adopting an ambiguity hypothesis: 'explains$_{\epsilon 1}$' is associated with an agent-theme relation, while 'explains$_{\epsilon 2}$' is associated with an agent-content relation. But other things equal, one wants to avoid such hypotheses and the meaning postulates that inevitably accompany them; and at least prima facie, the same matrix verb appears in (2) and (3). Alternatively, one might deny that (3) has the form described by (β). If x explained that Q, x did so *by* doing something else—for example, asserting that Q in response to a question. So perhaps the surface form of (3) masks a more complicated logical form like the following:

$$\exists e\exists x\exists f[\text{agent}(e, \text{Nora})\ \&\ \text{explaining}_\epsilon(e)\ \&\ \text{theme}(e, x)\ \&\ \text{by}(e, f)\ \&\ \text{agent}(f, \text{Nora})\ \&\ \text{asserting}(f)\ \&\ \text{content}(f, \text{that Fido barked})]$$

On this view, 'that Fido barked' is an argument of an unspoken verb, which belongs to an adverbial phrase. But one needs independent evidence for the unspoken material, and an account of why:

(15) Nora explained that Fido barked on Tuesday.

is not *three* ways ambiguous. Prima facie, the adverbial phrase 'on Tuesday' should be able to modify all three verbs: 'explain', 'bark', and the unvoiced 'assert'. But (15) cannot mean that Nora explained something by *asserting on Tuesday* that Fido barked. Nor can:

(16) On Monday, Nora explained on Tuesday that Fido barked.

mean that: Nora explained something on Monday, by asserting on Tuesday that Fido barked.[6]

If we reject (β), this has implications for sentences like:

(17) Nora said that Fido barked.
(18) Nora believes that Fido barked.

whose logical forms would be:

(17L) $\exists e$[agent(e, Nora) & saying(e) & content(e, that Fido barked)]
(18L) $\exists s$[subject(s, Nora) & believing(s) & content(s, that Fido barked)]

If (3) and (17) have the same logical form, we should pause before claiming that (17) is true if and only if Nora bears the saying-relation to the referent of 'that Fido barked'. Similarly, if (18) is true, Nora is the subject of a state *whose content is* that Fido barked. But if 'believe' is not true of subject-theme pairs, we need not think of believing as something *done to* a proposition.[7] (Compare 'grasp the proposition', which suggests a theme of the grasping.) So we should be wary of treating (17-18) on the model of:

(19) She kicked it.

For (17-19) may not be instances of the common logical form 'Rxy', unless 'R' stands for the (disjunctive) relation: x is the agent-or-subject of an event-or-state whose theme-or-content is y.

Fregeans should welcome this consequence: believing that P need not be construed as "mentally grabbing" and affirming a thought. But paratactic theorists face a serious difficulty. According to Davidson, the truth conditions of (17) are given by the second sentence in

---

[6] Cf. Jerry Fodor, "Three Reasons for Not Deriving 'Kill' from 'Cause to Die'," *Linguistic Inquiry*, I (1970): 429-38.

[7] In this regard, consider "measurement analogies" for belief ascriptions—as discussed in Robert Matthews, "The Measure of Mind," *Mind*, CIII (1994): 131-46.

(17P) Fido barked. Nora said that.

Here, 'that' is used to demonstrate (an utterance of) the first sentence.[8] One might extend this idea to

(1) Nora explained why Fido barked.
(1P) Fido barked. Nora explained why.

For the second sentence in (1P) arguably captures the truth conditions of (1), so long as the reference of 'why' in (1P) is fixed by the first sentence in (1P). But this analogy to (17) breaks down with respect to

(3). Nora explained that Fido barked.
(3P) Fido barked. Nora explained that.

For (3P) captures the truth conditions of (1), which is true if and only if Nora's *explanandum* was (the fact) that Fido barked; (3P) does not capture the truth conditions of (3), which is true if and only if Nora's *explanans* was (the fact) that Fido barked.

Correlatively, the demonstrative in (3P) is associated with a theme. The second sentence in (3P) is true if and only if $\exists e$[agent(e, Nora) & explaining(e) & theme(e, Δ)]; where Δ is the item demonstrated with 'that'. For Davidson, Δ will be an utterance of 'Fido barked'; although a paratactic theorist willing to posit facts—if only to avoid the consequence that explananda are *utterances*—could allow that Δ is the fact that Fido barked. But on any view, if the second sentence in (3P) is true, the referent of 'that' is the thing explained (and not the content of Nora's speech act). Hence, the second sentence in (3P) has *different* truth conditions than (3), in which 'that Fido barked' is associated with a content. In short, paratactic theorists need to express the truth conditions of (3) as follows: $\exists e$[agent(e, Nora) & explaining(e) & content(e, Δ)]. But this is just what (3P) fails to deliver.

III. CONCLUSION

The logical forms of (2) and (3) are, respectively, (2L) and (3L):

(2) Nora explained the fact that Fido barked.

---

[8] Donald Davidson, "On Saying That," *Synthese*, XIX (1968): 130-46. For more recent discussion, see Herman Cappelen and Ernest Lepore, "Varieties of Quotation," *Mind*, CVI (1997): 429-50.

(2L) $\exists e$[agent(e, Nora) & explaining(e) & theme(e, the fact that Fido barked)]

(3) Nora explained that Fido barked.

(3L) $\exists e$[agent(e, Nora) & explaining(e) & content(e, that Fido barked)]

The semantics of 'that'-clauses should be pursued within a thematically elaborated event analysis that makes a distinction between themes and contents. And this suggests that it may be time to retire:

> (β) Instances of 'x V that Q' are true, if and only if x bears the relation expressed by (the verb) 'V' to the referent of 'that Q'.

PAUL M. PIETROSKI

University of Maryland

COMMENTS AND CRITICISM

MEASURING CONFIRMATION AND EVIDENCE[*]

Bayesian epistemology suggests various ways of measuring the support that a piece of evidence provides a hypothesis. Such measures are defined in terms of a subjective probability assignment, pr, over propositions entertained by an agent. The most standard measure (where 'H' stands for "hypothesis" and 'E' stands for "evidence") is:

the difference measure: $d(H,E) = \mathrm{pr}(H/E) - \mathrm{pr}(H)$[1]

This may be called a "positive (probabilistic) relevance measure" of confirmation, since, according to it, a piece of evidence E *qualitatively* confirms a hypothesis H if and only if $\mathrm{pr}(H/E) > \mathrm{pr}(H)$, where qualitative disconfirmation is characterized by replacing '$>$' with '$<$' and confirmational irrelevance is characterized by replacing the '$>$' with '$=$'. Other more or less standard positive relevance measures that have been proposed are:

the log-ratio measure: $r(H,E) = \log[\mathrm{pr}(H/E)/\mathrm{pr}(H)]$

and

the log-likelihood-ratio measure: $l(H,E) = \log[\mathrm{pr}(E/H)/\mathrm{pr}(E/{\sim}H)]$[2]

---

[*] We thank Marty Barrett, Mike Byrd, Malcolm Forster, Dan Hausman, Ilkka Kieseppa, and Elliott Sober for useful comments.

[1] While a one-place subjective probability assignment, $\mathrm{pr}(-)$, is a measure of an agent's degrees of belief in propositions, $\mathrm{pr}(-/-)$ is a two-place *conditional* probability function. $\mathrm{pr}(H/E)$, the probability of H conditional on E, is (usually) defined as $\mathrm{pr}(H\&E)/\mathrm{pr}(E)$. Roughly, according to Bayesian epistemology, the probability of H conditional on E is supposed to correspond to the degree of belief the agent would have in H were the agent to learn E.

[2] The latter two measures take the *log* of the ratios simply to normalize, so that, like the difference measure, positive and negative confirmation correspond to measures > 0 and < 0, respectively, and confirmational irrelevance corresponds to measure $= 0$. Advocates of r include Peter Milne, "*log*[$p(h/eb)/p(h/b)$] Is the One True Measure of Confirmation," *Philosophy of Science*, LXIII, 1 (1996): 21-26; and George N. Schlesinger, "Measuring Degrees of Confirmation," *Analysis*, LV, 3 (1995): 208-12. Advocates of l include I. J. Good, "Explicativity, Corroboration, and the Relative Odds of Hypotheses," in his *Good Thinking: The Foundations of Probability and Its Applications* (Minneapolis: Minnesota UP, 1983), pp. 149-70; and Good, "The Best Explicatum for Weight of Evidence," *Journal of Statistical Computation and Simulation*, XIX (1984): 294-99.

0022-362X/00/9712/663–72

In a recent paper, David Christensen[3] criticizes these measures, and as a partial solution to the criticisms he raises, suggests what he regards as an improved measure, which we call:

the normalized difference measure: $S(H,E) = [1/\mathrm{pr}(\sim E)][d(H,E)]$
$= \mathrm{pr}(H/E) - \mathrm{pr}(H/\sim E)$[4]

Of course, this measure also is a positive relevance measure.

Christensen's criticisms of the standard measures focus mainly on the difference measure, and he claims that the same alleged difficulties apply also to the other measures, including, ultimately, his own suggested measure, *S*; his ultimate conclusion is skepticism about the possibility of defining an adequate probabilistic measure of confirmation. The objections he raises to these measures involve mainly the so-called "problem of old evidence"[5]; the claim is that the standard measures do not fare well in certain situations he describes involving "old evidence," while *S* does better in these situations but is itself not ultimately satisfactory. In section I, we argue that the standard measures (in particular the difference measure) can accommodate the examples Christensen describes just fine—that he has not fully appreciated the versatility of the relevant part of a kind of Bayesian resolution (that he discusses) of the problem. In section II, we argue that Christensen's allegedly improved measure *S* suffers from special difficulties of its own, besides also being vulnerable to the alleged "old-evidence" difficulties he describes for the standard measures, and we conclude with reflections on Christensen's skeptical conclusion.

I. OLD EVIDENCE AND SUPPORT OF HYPOTHESES

The problem of old evidence, as usually formulated, arises in situations in which $\mathrm{pr}(E) = 1$. In this case, $\mathrm{pr}(H/E) = \mathrm{pr}(H)$, and *E* cannot confirm *H*—on *any* of the *first three* measures of confirmation described above. But this can happen even when, intuitively, evidence *E* would seem to provide significant support for hypothesis *H*: when *E* was learned should not affect its evidential relevance for *H*.

---

[3] "Measuring Confirmation," this JOURNAL, XCVI, 9 (September 1999): 437-61.

[4] This gives two formulations of measure *S*, which Christensen proves are equivalent to each other (provided that $\mathrm{pr}(E) \neq 1$).

[5] This problem for Bayesian confirmation theory, to be discussed below, was, famously, raised by Clark Glymour in his *Theory and Evidence* (Princeton: University Press, 1980) and was noticed as an issue for probabilistic theories of confirmation as early as 1968: see Good, "Corroberation, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor," *The British Journal for the Philosophy of Science*, XIX (1968): 123-43; also his "A Historical Comment concerning Novel Confirmation," *The British Journal for the Philosophy of Science*, XXXVI, 2 (June 1985): 184-86.

Christensen says that the problem of old evidence, or at least a version of it, can arise even when pr(E) is not equal to 1, but just very close to 1; this is significant since many Bayesians feel uncomfortable about the idea that a rational person should ever assign probability 1 to a nontautology anyway. Christensen claims that "it remains true that as pr(E) approaches 1, the degree to which E can confirm anything becomes [on measures d and r] vanishingly small" (439).[6] Christensen emphasizes the version of the problem of old evidence in which pr(E) is not equal to 1. And he points out that in this case $S(H,E)$ need not vanish as pr(E) approaches 1.[7] But since the "old-evidence" criticisms he advances against the standard measures of confirmation do not depend on whether pr(E) equals or is only very close to 1, we shall, for simplicity, assume in this section the usual version in which pr(E) $= 1$. Further, we limit ourselves in the rest of this section to discussing the criticisms of the difference measure d. As Christensen says about his criticisms, our points also can easily be applied to the log-ratio measure r; the log-likelihood-ratio measure l will be a topic in section II.

Even given this, there are still two problems of old evidence that Christensen distinguishes, which he calls the *diachronic* and *synchronic* problems.[8] The diachronic problem involves an actual event of confirmation, at an earlier time at which evidence is learned and confirmation of a hypothesis actually takes place.[9] Suppose that the diachronic problem can be solved by showing that the relevant confirmation event can be modeled as follows: at the time of the confirmation event, the probability of the hypothesis H increases appropriately from its old unconditional probability to a new value equal to the old probability of *H* conditional on *E*. Given this, Christensen focuses on the synchronic problem of explaining, in Bayesian terms, how *E* can *still* be regarded as providing appropriately significant support for *H*, even though pr(*E*) is now equal to 1. As he puts it:

---

[6] Note that pr(H) = pr(H/E)pr(E) + pr($H/{\sim}E$)pr(${\sim}E$), so that pr(H) must approach pr(H/E) as pr(E) approaches 1. But this is consistent with the possibility that, for example, each of these two values fluctuate together from very close to 0 to very close to 1 as pr(E) approaches 1. But a natural and standard assumption (for example, as in Richard Jeffrey's probability kinematics, *The Logic of Decision* (Chicago: University Press, 1983)) is that pr(H/E) and pr($H/{\sim}E$) remain constant as the probabilities of the evidence propositions, E and ${\sim}E$, may vary. In this case, as pr(E) approaches 1, pr(H) will approach the fixed value pr(H/E); and, of course, the degree to which E supports H vanishes to 0, on each of the first three measures of support described above.

[7] In section II, we discuss this and other comparisons between S on the one hand and d, r, and l on the other.

[8] For other ways of classifying problems of old evidence, see the references cited in Christensen.

[9] There are actually several versions of the diachronic problem; Christensen's concern is the simple kind of problem where E can be thought of as an observation report. For discussion of two other kinds of diachronic problems, not relevant to his specific criticisms, see Christensen and the references cited therein.

> ...it seems clear that two things can be true simultaneously: (1) some beliefs will provide significant evidential support for others, and (2) some of the evidential beliefs will be held with high levels of confidence. These two features are...flatly incompatible on the standard Bayesian analysis of (synchronic) evidential support (443).

Christensen considers the following "historical approach" to the synchronic problem: "it seems appropriate to say that *E is (actual) evidence for H*, for a given individual, if, at some time in the past, the event of its confirming *H*, for that individual, took place."[10] Crucial here is the idea of distinguishing between the *past event of E's confirming H* from *E's now being part of one's current body of evidence for H*, where the choice of the terms 'confirming' and 'evidence' is somewhat arbitrary and meant just to mark the distinction. We have used the term 'support' above in some places to be neutral between these two distinct but related ideas.

These two ideas (especially the evidence idea) will be elaborated further and improved upon below to deal with the following objection raised by Christensen (444-45). He describes an example in which the hypothesis (call it *H*) is that deer live in a nearby wood. The agent comes across (call this evidence *D*) a pile of deer droppings there. The probability of *D* becomes 1 and the probability of *H* increases to near 1. Subsequently, he comes upon (call this evidence *A*) a shed deer antler. At this time the probability of *H* is already very high, however, so the discovery of *A* could not significantly increase the probability of *H*. Thus, on the "historical approach" described above, *D* provides significant support for *H* while *A* provides very little support. Intuitively, however, it would seem that *D* and *A* are equally strong evidence for *H*.[11] And as Christensen puts it:

> The historical approach thus makes contemporary evidential support depend in an unintuitive way on the order in which evidence was

---

[10] From Ellery Eells, "Problems of Old Evidence," *Pacific Philosophical Quarterly*, LXVI (1985): 283-302, here p. 287; we have changed here the '*T*'s in the original to '*H*'s.

[11] Approach of pr(*H*) to the particular value 1 is not crucial to the problem Christensen sees here. All that is important is that two pieces of evidence seem symmetrical in their support of *H* while in the presence of either the other does not affect the probability of *H* much.

> discovered. Intuitively, synchronic support should depend on the agent's present epistemic state, not on such historical accidents. Clearly, the synchronic problem cannot be reduced to the diachronic one in this way (445).

To introduce some notation that will be useful in describing the Bayesian solution to the synchronic problem which answers Christensen's criticisms, we first describe in a more general way the solution to the diachronic problem. Let E_1, E_2,..., E_N be all the evidence propositions which have been learned to date which are relevant to the hypothesis H; say these pieces of evidence came in at times t_1, t_2,..., t_N, respectively (so that it is now after t_N); let pr_0 be the subjective probability assignment the agent begins with, just before t_1; and, for $i = 1, 2,..., N$, let $\mathrm{pr}_i(-) = \mathrm{pr}_0(-/E_1\&E_2\&...\&E_i)$.[12] Now we can define *confirmation* as follows:

> Definition: *E_i confirms H at time t_i if and only if $pr_i(H)$* ($= \mathrm{pr}_{i-1}(H/E_i)$) $>$ *$pr_{i-1}(H)$;* and the *degree to which E_i confirms H at t_i is the difference between these two probabilities.*

Then, in Christensen's deer example, D significantly confirmed H at the time it was learned, but A did *not significantly* confirm H at the time *it* was learned, when D had become part of "background knowledge." Admittedly, this registers what may be a "historical accident" of the order in which the evidence came in. But this will seem unintuitive only if we leave out the rest of the Bayesian story, which uses the concept we call *being evidence for*, to which we now turn.

On the Bayesian account, hypothesis support is a three-place relation among evidence, hypothesis, and background knowledge (where the last is "summarized" in a subjective probability function which changes as evidence comes in). This shows up in the definition of confirmation above; and relativity to (different parts of) background knowledge is central to the definition of 'being evidence for', given below. As a simple application of the general definition to follow, we turn again to Christensen's deer example. Here we may say that relative to *no relevant background knowledge*, each of D and A is significant evidence for H, in that: $\mathrm{pr}_0(H/D) > \mathrm{pr}_0(H)$ and $\mathrm{pr}_0(H/A) > \mathrm{pr}_0(H)$, where in each case the difference is "significant." The relativity to the state of no relevant background knowledge is represented in the probability comparisons here by using, on the right-hand sides, the pr_0-unconditional probability of H, and on the left-hand sides, that probability conditional on the evidence. Also, relative to *background knowledge D (A), A (D)* is something less than "significant" evidence for H, in that $\text{pr}_0(H/D\&A)$ ($\text{pr}_0(H/A\&D)$) is only slightly larger than $\text{pr}_0(H/D)$ ($\text{pr}_0(H/A)$). The relativity to background knowledge D (A) is represented here in the probability comparisons by taking both probabilities compared to be conditional on D (A). Note that here the question of the significance of evidence for hypothesis does not depend on the order in which evidence comes in.

---

[12] pr_0 should *not* be understood as representing an agent's degrees of beliefs before she has encountered *any* evidence about *anything at all*, or a "pure" initial probability function with which an agent enters the world. Rather, we are simply backing up to the point at which the agent lacks the evidence relevant to the analysis of a particular case of hypothesis support. (Compare Christensen, 460, footnote 34.)

For the sake of completeness, we now give a more general characterization of *evidence*.[13] Where again $E_1,..., E_N$ includes all the evidence that is relevant to a hypothesis H, and pr_0 is as before, we can ask what the evidential relevance of some E_i is for H relative to some subset of the propositions $E_1,..., E_{i-1}, E_{i+1},..., E_N$. Let B range over the 2^{N-1} conjunctions of the members of these 2^{N-1} subsets. Then we offer the following definition:

> Definition: *E_i is evidence for H, relative to B,* if and only if $\text{pr}_0(H/B\&E_i) > \text{pr}_0(H/B)$; and the *degree of E_i's evidential support for H, relative to B,* is the difference between these two probabilities.

We emphasize that B can be the conjunction of the members of *any* subset of $\{E_1,..., E_{i-1}, E_{i+1},..., E_N\}$ and not just of the set $\{E_1,..., E_{i-1}, E_{i+1},..., E_N\}$ itself.

Making use of the idea that hypothesis support is a three-place relation, we have characterized two distinct questions that can be asked about support. One can ask whether E actually confirmed H, and this question is implicitly relative to the actual history of the relevant subjective probability assignment, in particular to the actual order in which relevant pieces of evidence came in. On the other hand, one can ask whether evidence E is (still and permanently) part of the evidence for H relative to some background knowledge B, and of course the answer will vary as we vary B. The first question is historical and depends on historical accident pertaining to the order in which evidence actually came in. But we agree with Christensen that there is an aspect of support that is more like a logical relation, being insensitive to historical accident, and permanent. We think these distinct ideas are captured by the two ideas we have defined and have called *confirmation* and *evidence.*

---

[13] This is an improvement over Eells (*op. cit.*, pp. 286-87); see also Eells, "Bayesian Problems of Old Evidence," in C. Wade Savage, ed., *Scientific Theories* (Minneapolis: Minnesota UP, 1990), pp. 208-10.

It is worth briefly mentioning another "problem" Christensen sees for the measure d (and all this applies equally to r), which he calls the *probable-hypothesis problem* (448-49). As he correctly points out, as $\mathrm{pr}(H)$ approaches 1, the difference between $\mathrm{pr}(H/E)$ and $\mathrm{pr}(H)$ must vanish to 0. His complaint is that even though E may, intuitively, be highly significant for H, the measure d can in this case measure only the (vanishingly small) historical impact of E on H. We agree that this is true, but appropriately so for our historical concept of confirmation, while it is also easy to see that, again, the approach to evidence detailed above does not put this kind of constraint on the "evidential" significance of E for H, as measured by d relative to various bodies of background knowledge B.

II. AN ASSESSMENT OF THE MEASURE S AND OF CHRISTENSEN'S SKEPTICAL CONCLUSION

In the end, Christensen is skeptical about the possibility of *any* adequate probabilistic measure of support, and offers his measure S just as an *improvement* over the standard measures he considers. In section I, we argued that Christensen's "old-evidence" criticisms of measure d (and of r) do not point to an advantage of S, when d (or r) is properly applied to the old-evidence issue. Now we consider (1) his assessment of the measure l, (2) his "normalization" idea (dividing a measure by $\mathrm{pr}(\sim E)$) in general, and (3) his reasons for his skeptical conclusion. And we conclude (a) that the normalized difference measure S (and normalization in general) does not represent an improvement over any of the standard measures, and (b) that Christensen's reasons for his skeptical conclusion are misplaced, in a hope for an "all purpose," "purely synchronic" measure of support (which seems to have been a motivation for S).

Christensen (440) is careful to point out that l, like S, does not violate what he takes to be two desiderata of measures c of support, discussed above, namely,

D_1: $\mathrm{pr}(E)$'s approaching 1 *does not imply* that $c(H,E)$ approaches 0

and

D_2: $\mathrm{pr}(H)$'s approaching 1 *does not imply* that $c(H,E)$ approaches 0

These are violated by d and r. He seems to suggest (440, and see his footnote 8), however, the following as a desideratum (a sort of variation on D_1), which is violated by l but not by S:

D_3: Provided that $\mathrm{pr}(H)$ does not (at the same time) approach 1, $\mathrm{pr}(E)$'s approaching 1 *does not imply* that $c(H,E)$ approaches 0.

Apparently, it is on this basis that Christensen thinks we should favor *S* over *l*. We think this criticism of *l* is too hasty—at least as a defense *of S* as compared to the log-likelihood-ratio idea. For one thing, it seems natural for defenders of *l* to ask why they should not also be allowed to use the trick of multiplying their measure by the "normalization factor" $1/\mathrm{pr}(\sim E)$. This would yield the following "normalized" version of the log-likelihood-ratio measure *l*:

$$l_N(H,E) = [1/pr(\sim E)]\, log[pr(E/H)/pr(E/\sim H)] = [1/pr(\sim E)][l(H,E)]$$

It is interesting to note that l_N satisfies all three of Christensen's desiderata D_1-D_3. In addition, it turns out that l_N is *even more* insensitive to the kinds of changes in $\mathrm{pr}(H)$ and $\mathrm{pr}(E)$ that seem to worry Christensen. In particular, l_N satisfies the following desideratum, but *S* does not:

> D_4: pr(H)'s approaching 0 while (at the same time) $\mathrm{pr}(E)$ approaches 1 *does not imply* that $c(H,E)$ approaches 0

For instance, if $\mathrm{pr}(H) = 0.01$ and $\mathrm{pr}(E) = 0.99$, then $S(H, E) \leq 0.0101$. By contrast, in such a case, there is nothing preventing $l_N(H,E)$ from being quite high. It seems intuitive that in cases where $\mathrm{pr}(H)$ is low it should be *easy* for *E* to provide significant support to *H* (even if $\mathrm{pr}(E)$ also happens to be high). According to Christensen's measure *S*, however, this cannot happen. It seems to us that Christensen should really be comparing his normalized measure *S* with other normalized measures like l_N, not with *non*normalized measures like *l*. After all, Christensen does not give up on the difference measure entirely—he just normalizes it so that it will satisfy certain desiderata.

On the other hand, one must consider the *costs* of normalization. For example, many contemporary Bayesian resolutions of both the ravens paradox and the problem of evidential variety (or diversity) depend on the following assumption about Bayesian measures of confirmation *c*:[14]

> D_5: If $\mathrm{pr}(H/E_1) > \mathrm{pr}(H/E_2)$, then $c(H, E_1) > c(H, E_2)$

It is not an exaggeration to say that most Bayesian confirmation theorists would accept D_5 as a *desideratum* for Bayesian measures of confirmation. Indeed, a wide variety of Bayesian relevance measures satisfy D_5, including *d*, *r*, and *l*. Unfortunately, if we "normalize" *any* of these three measures, we end up with a measure that violates D_5. In particular, *S* and l_N *both* violate D_5.[15]

---

[14] See, for instance, Paul Horwich's *Probability and Evidence* (New York: Cambridge, 1982) for a Bayesian explication of the ravens paradox which depends on D_5 (pp. 54-63) and for an outline of a Bayesian explication of the confirmational value of varied data which also depends on D_5 (pp. 118-22).

Finally, we turn to why Christensen himself ultimately rejects *S* and takes his skeptical position about the possibility of an adequate probabilistic measure of support. Turning back to his deer example, Christensen asks us to suppose that (in our notation) $\mathrm{pr}_0(H) = 0.5$, and $\mathrm{pr}_0(D)$ and $\mathrm{pr}_0(A)$ are both very low. At t_0, both *D* and *A* would confirm *H* very strongly (on any measure of confirmation we have considered). Then, at t_1, the agent discovers what are almost certainly deer droppings. So, both $\mathrm{pr}_1(H)$ and $\mathrm{pr}_1(D)$ are very high, and as Christensen notes, "*S*-support from *A* becomes very low" just after t_1. According to Christensen, "this is quite unintuitive; *A* seems just as good a sign of deer as it was before *D* became highly probable" (457). For this reason, Christensen concludes (459) that *S* gives the wrong answer in this example.

Christensen offers an explanation of why *S* goes wrong in this kind of example: "*S* is insensitive to...the distinction between specific evidence and background assumptions" (459). But rather than conclude from this interesting observation that choosing a particular mathematical form for a measure of hypothesis support will not by itself be enough to properly handle the problem of old evidence, Christensen ends up with the general, skeptical conclusion: "It now seems that probabilistic accounts will...miss our ordinary notion's dependence on the distinction between background beliefs and specific evidence" (461).

We suggest it is mistake to think that the synchronic old-evidence problem must be solved (if solvable at all) by "purely synchronic means"—that is, merely by fiddling with the mathematical forms of measures of confirmation, without recourse to different parts of the agent's background knowledge. The fact that such "purely synchronic" approaches to the problem of old evidence seem doomed to failure is no reason to be skeptical about *all* probabilistic approaches to the problem of old evidence. As we have suggested, when the theory of probabilistic hypothesis support is elaborated to include both a notion (and measure) of confirmation and a notion (and measure) of evidence that explicitly incorporate, in appropriately different ways, relativity of hypothesis support to parts of background knowledge, the problem of old evidence can be adequately resolved.

---

[15] See Branden Fitelson, "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity," *Philosophy of Science*, LXVI, 3 (Proceedings Supplement) (1999): S362-78, for further discussion of D_5, including a proof that the relevance measures of Rudolf Carnap [*Logical Foundations of Probability* (Chicago: University Press, 1962, 2nd ed.)], Robert Nozick [*Philosophical Explanations* (Cambridge: Harvard, 1981)], and Halina Mortimer [*The Logic of Induction* (Englewood Cliffs, NJ: Prentice Hall, 1988)] also violate D_5. For further unintuitive features of the measure *S*, see Eells's review of James M. Joyce's *The Foundations of Causal Decision Theory*, in *The British Journal for the Philosophy of Science*, LI, 4 (December 2000, forthcoming); Fitelson, "A Bayesian Account of Independent Evidence with Applications," *Philosophy of Science*, LXVIII (2001, forthcoming); and Eells and Fitelson, "Symmetries and Asymmetries in Evidential Support" (manuscript).

ELLERY EELLS
BRANDEN FITELSON

University of Wisconsin/Madison

BOOK REVIEWS

The Social Construction of What? IAN HACKING. Cambridge: Harvard University Press, 1999. x + 261 p. Cloth $29.95, paper $16.95.*

It is no secret that many mainstream philosophers of science have held much recent sociology of science in low esteem, and no doubt the feeling is sometimes mutual. At bottom, the attitude of such philosophers derives from the feeling that sociologists have not given science sufficient respect, respect especially for that cardinal virtue in the philosophical lexicon, truth. Sociologists, on the other hand, have often considered the question whether (some) scientific claims are true as shedding little or no light on the proper understanding of the actual processes through which those claims are evaluated and accepted as knowledge. Philosophers have traditionally been concerned with what, if anything, gives science a unique access to the truth about nature, while sociologists have been concerned rather with how scientists reach and defend the conclusions they assert, and the two groups have often had little to say to one another. While constituting a regrettable fault-line in the academic environment, this mutual disrespect was for many years hardly a conspicuous feature of the intellectual landscape.

Then, most notably with the publication of Paul Gross and Norman Levitt's *Higher Superstition: The Academic Left and Its Quarrels with Science,*[1] and subsequently with Alan Sokal's notorious spoof published in the journal *Social Text,*[2] a number of scientists joined battle very publicly against the sociologists and a miscellaneous assortment of radical philosophers and literary theorists, and the Science Wars had begun. Although a few philosophers weighed in on the side of these scientific warriors, on the whole even those who had been happy in the past to express their contempt for "social constructivist" sociologists were reluctant to align themselves with the crude polemics of Gross, Levitt, and Sokal. In the public eye, the debate was generally presented as a battle between the defenders of science and the attacks of loony left-wing academics trying to claim that science was simply made up by cynical scientists pursuing their own personal

---

* I am grateful to Barry Barnes and Adrian Haddock for helpful comments on an earlier draft.

[1] Baltimore: Johns Hopkins, 1994.

[2] "Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity," *Social Text*, 46/47 (Spring/Summer 1996): 217-52.

0022-362X/00/9712/673–76

aggrandizement. Richard Dawkins's often quoted remark that no one is a social constructionist at 30,000 feet nicely summarized the assumption that this antiscientific attitude was not one that a serious and responsible person could honestly maintain.

Issues are, of course, a lot more complex than this, and it is very timely that a philosopher of the breadth and subtlety of Ian Hacking has devoted himself to unraveling many of their complexities. The *Social Construction of What?* is, therefore, a very important book. As I have indicated, the crude statement of the debate opposes defenders of the truth or approximate truth of scientific results to social constructionists. One might well begin by worrying about why these positions should be in opposition at all. Naively, science is quite obviously constructed (by scientists, a social group) but some of it may perfectly well be true, too. No one thinks cars are unreal because they are constructed by groups of people. Although Hacking does much to problematize the crude opposition just mentioned, he does not try to dissolve it. Rather, according respect to protagonists on both sides, he offers a convincing diagnosis of what are the real philosophical issues separating them.

Hacking proposes three philosophical issues, or "sticking points," as central to the controversy between constructivists and realists. First, constructivists defend a contingency thesis: the actual development of science could have been quite different. Central to this idea is a kind of path dependence of scientific knowledge. Science at any time makes some questions interesting and others unintelligible. The development of particular methods of investigation or kinds of apparatus directs attention to some phenomena but not others. Although I have described the debate as opposing constructivists to realists, as a matter of fact Hacking thinks contingency is quite consistent with realism about science (79-80). Any of these paths through the domain of possible scientific knowledge may lead us to the discernment of real entities and perhaps also to the formulation of true propositions about them. One might even imagine that in the long run (when we shall, of course, all be dead) these different paths will all lead to the same destination. What we cannot suppose (if anyone was tempted to do so) is that what we now contingently believe is the whole truth about nature or, more controversially, the only possible truth about an aspect of nature. And though contingency may be consistent with scientific realism, science warriors on the realist side tend to deny contingency. Typically, they will say that Maxwell's equations or the second law of thermodynamics are necessary outcomes of any successful scientific history, which is what the contingency thesis is committed to denying.

Closer to a direct contradiction with contemporary realism is the second sticking point, nominalism. Constructivists are typically nominalists in the sense that they do not believe that there is any uniquely privileged inherent structure to the world that pre-exists our attempts to describe it. To invoke the inevitable figure of butchery, the most radical may hold that nature has no joints at all, the more moderate that she has no set of joints uniquely suitable for carving. And the third sticking point is the question of how the stability of scientific belief is explained. Hacking quotes Steven Weinberg[3] describing Maxwell's equations as having "settled down to a final form, valid in certain known circumstances...[I]n this form and in this limited context they have survived for a century and may be expected to survive indefinitely" (86). This remark nicely ties together the three sticking points. For Weinberg, scientists have discovered Maxwell's equations because they are an essential (noncontingent) part of any adequate physics, describing a basic part of the real structure of things (contra nominalism). That is why they will survive indefinitely. Constructivists will need a different account of whatever stability they believe to obtain. A well-known example of such an account is Bruno Latour's[4] description of the "network of agents and events that lie behind an item of knowledge" (90). Hacking himself is ambivalent about all of these theses, though largely favorable to nominalism, and somewhat hostile to contingency. He holds, at least for the natural sciences, that once questions have been sufficiently clearly formulated, nature will pretty much dictate the answers. This does, of course, leave a good deal of potential contingency about what questions are even asked.

Constructivism is not, however, merely a set of philosophical theses. Constructivists frequently have critical agenda in describing how particular parts of science are socially constructed. These range from merely historical and largely uncontroversial accounts of the actual history of a scientific idea, to the deployment of this history in exposing, or unmasking, the contingency of the ways of thought thus exposed and, finally, to attempting to overthrow these ways of thought. Obvious examples of the more extreme such agenda can be found in critical discussions of purportedly scientific accounts of race and gender. Often, though by no means always, social constructionists think that what has been constructed is a bad thing.

---

[3] "Sokal's Hoax," *New York Review of Books* (August 8, 1996): 11-15.

[4] *Science in Action: How to Follow Scientists and Engineers through Society* (Cambridge: Harvard, 1987).

Consideration of these critical functions of constructivism leads to perhaps the clearest, and perhaps the most controversial, distinction that Hacking defends. He suggests that there is a fairly sharp boundary between the social and the natural sciences, grounded in the fact that classifications in the social sciences have profound effects upon, and can in turn be affected by, the subjects of that classification. They are what he calls interactive or looping kinds. Here, many readers will recall Hacking's recent work on various mental-health epidemics, and in this volume is a chapter on the construction of the categories of child abuse, abusers, victims, and so on. The point of this Foucauldian vein in Hacking's work is not, of course, to show, for instance, that nobody mistreated children before 1961, but that the development of child abuse as a scientific concept had profound effects on the lives of countless children and their more or less culpable caretakers. And these effects, in turn, inevitably affected the truth of claims about child abuse. Although Hacking identifies important aspects of all the philosophical inclinations he identifies with constructivism in examples from the physical sciences, this kind of interaction, he insists, is peculiar to the classification of humans: only humans are aware of being classified and conform their behavior to that awareness. And, of course, this aspect of the construction of social science makes a clear, and sometimes critical, perspective on the process all the more important.

It is something of a cliché in reviewing to note that little of the richness of the work in question can be captured in a review of the allocated length. This applies with particular force to the present book. I have tried to summarize the main philosophical themes, but the real richness of the book is in the variety and depth of the examples with which these themes are illustrated. (These include: child abuse, hypothetical nanobacteria in dolomite, weapons research, the voyages of Captain Cook, and quarks.) In all of these domains, there is ample space for debate on each of the sticking points Hacking distinguishes. Traditional philosophy of science has much to tell us about these domains and so does social constructivism. But the details are quite different for these very disparate domains. Perhaps the most important moral, and one that depends directly on this rich variety of examples, is the diversity of science. Neither of these perspectives has the whole truth about science, because there is no such whole truth. Various stories can be constructed about particular parts of science, and many of them may well be true.

JOHN DUPRÉ

University of Exeter

Rethinking the Foundations of Statistics. JOSEPH B. KADANE, MARK J. SCHERVISH, and TEDDY SEIDENFELD. New York: Cambridge University Press, 1999. x + 388 p. Cloth $64.95, paper $24.95.

Since Leonard J. Savage published *The Foundations of Statistics*[1] in 1954, the face of statistics has changed. This is not to say that subjective Bayesianism has swept statistics like a storm, but rather that frequentist statistics—still by and large the dominant view—has become cognizant of new issues and new challenges. Philosophy (and to a large extent artificial intelligence), on the other hand, does seem to have been taken by storm. Subjective Bayesianism is so politically correct in philosophy that references to any alternative are often met with blank stares.

This volume, written from a loosely Bayesian point of view, shows clearly that "Bayesianism" is no monolithic and settled doctrine, but comes in many importantly different flavors, and that many of the open issues are of deep philosophical importance. Often, these issues arise in the context of decision making that involves more than a single Bayesian agent. Thus the first part of the volume concerns cooperative decision making, and the third part concerns noncooperative decision making, or game theory. The second (and longest) section of the book concerns somewhat more technical (but nonetheless philosophically important) issues, such as the questions raised by state-dependent utilities or countable (as opposed to finite) additivity.

The essays address six themes (4). Four concern multiagent decisions (though in fact some arise in the case of a single agent who is less than compulsively Bayesian):

(1) cooperative nonsequential decisions
(2) the representation of partially ordered preferences
(3) noncooperative sequential decisions (game theory)
(4) pooling rules and Bayesian dynamics for sets of probabilities

Two other themes concern:

(5) state-dependent utilities
(6) additivity assumptions (finite or countable)

The authors delineate, often in ways that are surprising, "the scope and limitations of current Bayesian Theory" (3). For example, the first essay addresses the question of whether the norms of expected-utility theory can be extended to a set of coherent agents subject to the weak Pareto rule: the group strictly prefers h_1 to h_2 provided every individual does. The answer is "no," but a weakening of the usual ordering assumption allows this.

---

[1] New York: Wiley, 1954.

The third essay develops a Bayesian theory based on partially ordered preferences. The motivation, again, is the desire to deal with cooperative decisions. One might also motivate this development by the desire to treat formally the "imperfect" Bayesian agent, whose preferences are incomplete. Indeed, this motivation could be taken quite broadly, and could apply to a number of the essays. It is surely clear that there is a close relation between the deliberations of n Bayesians, and the deliberation of a single Bayesian who is of n minds. Thus we may consider agent A who attributes a probability of 0.4 to the occurrence of heads on the next toss of a coin, and agent B who attributes a probability of 0.7 to the occurrence of heads on the next toss of the coin. These two agents will certainly agree about many alternatives—for example, what to do in response to an enthusiastic bettor who offers 10:1 odds against heads!

But now consider a single agent who is sure that the coin comes from a bag of 40% coins or a bag of 70% coins, but who is *not a proper Bayesian,* in the sense that he will assign no value to the probability that the coin comes from the 40% bag. He is of two minds, and the two minds admit of no mixture. Still, there are a number of preferences that he clearly has, and, like the agents A and B, his opinions can be influenced by evidence.

Of course, we can *impose* convexity on the beliefs of the agent in question: that is, we can require that, if the agent is willing to bet at odds of 4 to 6, or at 7 to 3, then he "should" also be willing to bet at odds of n to m, where $n/(n + m) \in [0.4,0.7]$. Such, for example, is Isaac Levi's view.

The situations are not exactly parallel, however. It is not plausible to impose convexity on the agents A and B. On the other hand, convexity imposed on the single agent might be taken to imply that the agent is of an infinite number of minds, or corresponds to an infinite community. Formally, this is not so: either convexity, *or the lack of it,* among agents can be represented within the framework used by the authors of these essays, which is to their credit. Nevertheless, it is disappointing that the analogy between two agents with one mind apiece, and one agent of two minds has its limitations.

In keeping with the general multiagent theme of the volume, the essay "Randomization in a Bayesian Perspective" finds the justification of randomization in statistical inference in the needs of *groups* of agents: randomization will allow groups of Bayesian agents to come to agreement despite their possibly different initial distributions. This will not strike a frequentist statistician as a strong enough conclusion. The frequentist will see a use for randomization even for a single agent. The authors conclude that when only one decision maker is relevant, randomization is a mistake, since it discards information. So, although a defense of randomization is provided, it is not a defense that would be regarded as sufficient by all statisticians (or even all philosophers). Even some Bayesians (for example, I. J. Good) have offered stronger justifications for randomization.

The issue of randomization is particularly controversial in the context of clinical trials, a topic that is treated briefly in the last section of this same essay. Random assignment of patients to treatments is a feature of many clinical trials. But it can be argued from an ethical point of view that each patient should receive that treatment which maximizes his expected utility, and, from a Bayesian point of view, these expected utilities always exist. Alleviating the ethical tension between randomization, allegedly required for statistical soundness, and doing the best thing for each individual patient are the subject of considerable discussion and research. The references provided in this essay can serve as an introduction to that tantalizing ethical literature.

Another topic that will be familiar to many is the following issue. According to Bayesian doctrine, it does not matter why you stop sampling; so it is possible to sample until the evidence supports your favorite hypothesis, and then stop. The essay, "Reasoning to a Foregone Conclusion," explores this topic. It turns out that finitely additive probabilities (corresponding to the use of improper priors) do allow arguing to a foregone conclusion! Various constraints on prior distributions have been proposed which will avoid the possibility of arguing to a foregone conclusion; the authors discuss them, but conclude only that "Whether in each case [the cost of foregoing the use of improper priors] is too high is a question for further investigation and serious debate" (367).

There are enough formulas and mathematical technicalities to challenge the most numerate of us, and yet the ratio of clear and literate prose to formulas is very high. Furthermore, in six of the sixteen essays, the technicalities are quarantined in appendices. Although the essays are, as required by their focus, technical, the issues are very general, and impinge on topics in logic, ethics, and the philosophy of science. Furthermore, they are often stated clearly and dramatically, and in simple intuitive terms.

The editors (who are also, with three additional co-authors, the authors of the essays) are committed "Bayesians"; two of them are thoughtful and sophisticated statisticians, and the third, coauthor of all but three of the essays, is a well-known philosopher of science. The essays are anything but doctrinaire, and in fact one needed lesson that emerges clearly from the essays is that there is no single Bayesian doctrine. There are choices to be made and issues to be confronted, even within the broad general outlines of a Bayesian approach. Although the authors of these essays are Bayesians, there is valuable material here for everyone interested in decision theory, game theory, or the foundations of probability and statistics.

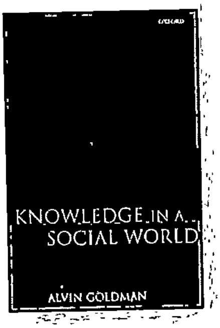
HENRY E. KYBURG, JR.

University of Rochester

NEW FROM OXFORD

KNOWLEDGE IN A SOCIAL WORLD

ALVIN I. GOLDMAN



"Goldman's book [shows] breadth of vision, mastery of relevant contemporary thought...practical utility, and...theoretical depth. It is a work crafted with supreme philosophical skill" —*SOCIAL EPISTEMOLOGY*

"This is a book of singular importance to lawyers, political theorists, social and natural scientists as well as to educators and theorists of education.... It may well be the most significant interdisciplinary philosophy book of the decade."—JULES COLEMAN, YALE LAW SCHOOL.

Against the tides of postmodernism and social constructionism Goldman defends the integrity of truth and shows how to promote it by well-designed forms of social interaction. The result is a bold, timely, and systematic treatment of the philosophical foundations of an information society.

1999 424 pp.; 3 line illus paper $19.95 cloth $70.00

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our website at www.oup.com

INDEX TO VOLUME XCVII

INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

AGM

Rott, Hans: Two Dogmas of Belief Revision. **97** (9): 503-22 '00

Bayesian Theory

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Belief

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Gendler, Tamar Szabó: The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

Rott, Hans: Two Dogmas of Belief Revision. **97** (9): 503-22 '00

Schick, Frederic: Surprise, Self-Knowledge, and Commonality. **97** (8): 440-53 '00

Biology, Philosophy of

Sober, Elliott: Evolution and the Problem of Other Minds. **97** (7): 365-86 '00

Causation

Collins, John: Preemptive Prevention. **97** (4): 223-34 '00

Hall, Ned: Causation and the Price of Transitivity. **97** (4): 198-222 '00

Lewis, David: Causation as Influence. **97** (4): 182-97 '00

Paul, L. A.: Aspect Causation. **97** (4): 235-56 '00

Schaffer, Jonathan: Trumping Preemption. **97** (4): 165-81 '00

Christensen, David

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Church, Alonzo

Copeland, B. Jack: Narrow versus Wide Mechanism: Including a Reexamination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Cognitive Science

Abbott, Barbara: Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

Copeland, B. Jack: Narrow versus Wide Mechanism: Including a Reexamination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Coherence

Millgram, Elijah: Coherence: The Price of the Ticket. **97** (2): 82-93 '00

Consequentialism

Cocking, Dean: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Kennett, Jeanette: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Sen, Amartya: Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00

Contractualism: *See* Contractarianism.

Contractariansim

Ashford, Elizabeth: Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Stark, Cynthia A.: Hypothetical Consent and Justification. **97** (6): 313-34 '00

Counterfactuals

Collins, John: Preemptive Prevention. **97** (4): 223-34 '00

Hall, Ned: Causation and the Price of Transitivity. **97** (4): 198-222 '00

Lewis, David: Causation as Influence. **97** (4): 182-97 '00

Paul, L. A.: Aspect Causation. **97** (4): 235-56 '00

Schaffer, Jonathan: Trumping Preemption. **97** (4): 165-81 '00

Epistemology: *See* Knowledge, Theory of.

Ethics

Ashford, Elizabeth: Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Cocking, Dean: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Kennett, Jeanette: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Nelkin, Dana K.: Two Standpoints and the Belief in Freedom. **97** (10): 564-76 '00

Sen, Amartya: Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00

Stark, Cynthia A.: Hypothetical Consent and Justification. **97** (6): 313-34 '00

Ethics, Kantian

Cocking, Dean: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Kennett, Jeanette: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Nelkin, Dana K.: Two Standpoints and the Belief in Freedom. **97** (10): 564-76 '00

Evolution

Sober, Elliott: Evolution and the Problem of Other Minds. **97** (7): 365-86 '00

Explanation

Collins, John: Preemptive Prevention. **97** (4): 223-34 '00

Hall, Ned: Causation and the Price of Transitivity. **97** (4): 198-222 '00

Lewis, David: Causation as Influence. **97** (4): 182-97 '00

Paul, L. A.: Aspect Causation. **97** (4): 235-56 '00

Pietroski, Paul M.: On Explaining That. **97** (12): 655-62 '00

Schaffer, Jonathan: Trumping Preemption. **97** (4): 165-81 '00

Fiction

Gendler, Tamar Szabó: The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

Fodor, Jerry A.

Abbott, Barbara: Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

Freedom

Nelkin, Dana K.: Two Standpoints and the Belief in Freedom. **97** (10): 564-76 '00

Frege, Gottlob

García-Carpintero, Manuel: A Presuppositional Account of Reference Fixing. **97** (3): 109-47 '00

Friendship

Cocking, Dean: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Kennett, Jeanette: Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Gödel, Kurt

Floyd, Juliet: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Gaifman, Haim: What Gödel's Incompleteness Result Does and Does Not Show. **97** (8): 462-70 '00

George, Alexander: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Putnam, Hilary: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Tennant, Neil: Deductive versus Expressive Power: A Pre-Gödelian Predicament. **97** (5): 257-77 '00

Velleman, Daniel J.: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Knowledge, Theory of

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Faulkner, Paul: The Social Character of Testimonial Knowledge. **97** (11): 581-601 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Millgram, Elijah: Coherence: The Price of the Ticket. **97** (2): 82-93 '00

Rott, Hans: Two Dogmas of Belief Revision. **97** (9): 503-22 '00

Schick, Frederic: Surprise, Self-Knowledge, and Commonality. **97** (8): 440-53 '00

Stoljar, Daniel: Physicalism and the Necessary a Posteriori. **97** (1): 33-54 '00

Vogel, Jonathan: Reliabilism Leveled. **97** (11): 602-23 '00

Wright, Crispin: Truth as Sort of Epistemic: Putnam's Peregrinations. **97** (6): 335-64 '00

Kripke, Saul

Stoljar, Daniel: Physicalism and the Necessary a Posteriori. **97** (1): 33-54 '00

Imagination

Gendler, Tamar Szabó: The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

Integrity

Ashford, Elizabeth: Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Justification

Stark, Cynthia A.: Hypothetical Consent and Justification. **97** (6): 313-34 '00

Language, Philosophy of

Abbott, Barbara: Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

Cameron, J. R.: Numbers as Types. **97** (10): 529-63 '00

García-Carpintero, Manuel: A Presuppositional Account of Reference Fixing. **97** (3): 109-47 '00

Lepore, Ernie

Abbott, Barbara: Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

Lewis, David

Collins, John: Preemptive Prevention. **97** (4): 223-34 '00

Hall, Ned: Causation and the Price of Transitivity. **97** (4): 198-222 '00

Paul, L. A.: Aspect Causation. **97** (4): 235-56 '00

Schaffer, Jonathan: Trumping Preemption. **97** (4): 165-81 '00

Logic, Philosophy of

Tennant, Neil: Deductive versus Expressive Power: A Pre-Gödelian Predicament. **97** (5): 257-77 '00

Make Believe

Gendler, Tamar Szabó: The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

Mathematics, Philosophy of

Cameron, J. R.: Numbers as Types. **97** (10): 529-63 '00

Floyd, Juliet: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Putnam, Hilary: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Tennant, Neil: Deductive versus Expressive Power: A Pre-Gödelian Predicament. **97** (5): 257-77 '00

McCall, Storrs

Gaifman, Haim: What Gödel's Incompleteness Result Does and Does Not Show. **97** (8): 462-70 '00

George, Alexander: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Velleman, Daniel J.: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Memo to Authors

97 (2): 107 '00

Metaphysics

Collins, John: Preemptive Prevention. **97** (4): 223-34 '00

Hall, Ned: Causation and the Price of Transitivity. **97** (4): 198-222 '00

Lewis, David: Causation as Influence. **97** (4): 182-97 '00

Paul, L. A.: Aspect Causation. **97** (4): 235-56 '00

Schaffer, Jonathan: Trumping Preemption. **97** (4): 165-81 '00

Tennant, Neil: Deductive versus Expressive Power: A Pre-Gödelian Predicament. **97** (5): 257-77 '00

Mind, Philosophy of

Copeland, B. Jack: Narrow versus Wide Mechanism: Including a Reexamination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Shapiro, Lawrence A.: Multiple Realizations. **97** (12): 635-54 '00

Sober, Elliott: Evolution and the Problem of Other Minds. **97** (7): 365-86 '00

Stoljar, Daniel: Physicalism and the Necessary a Posteriori. **97** (1): 33-54 '00

Multiple Realizability Thesis

Shapiro, Lawrence A.: Multiple Realizations. **97** (12): 635-54 '00

New Books

97 (7): 409-16 '00

New Books: Anthologies

97 (7): 417-20 '00

97 (9): 528 '00

New Books: Paperback Reprints

97 (11): 633-34 '00

New Books: Translations

97 (8): 476 '00

Notes and News

97 (2): 107 '00

Numbers

Cameron, J. R.: Numbers as Types. **97** (10): 529-63 '00

Obligations

Sen, Amartya: Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00

Philosophy, Political: *See* Political Theory.

Physicalism

Stoljar, Daniel: Physicalism and the Necessary a Posteriori. **97** (1): 33-54 '00

Political Theory

Stark, Cynthia A.: Hypothetical Consent and Justification. **97** (6): 313-34 '00

Practical Reasoning

Sen, Amartya: Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00

Probability

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Psychology, Philosophy of

Copeland, B. Jack: Narrow versus Wide Mechanism: Including a Reexamination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Gendler, Tamar Szabó: The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

Shapiro, Lawrence A.: Multiple Realizations. **97** (12): 635-54 '00

Sober, Elliott: Evolution and the Problem of Other Minds. **97** (7): 365-86 '00

Putnam, Hilary

Shapiro, Lawrence A.: Multiple Realizations. **97** (12): 635-54 '00

Wright, Crispin: Truth as Sort of Epistemic: Putnam's Peregrinations. **97** (6): 335-64 '00

Rationality

Nelkin, Dana K.: Two Standpoints and the Belief in Freedom. **97** (10): 564-76 '00

Rawls, John

Stark, Cynthia A.: Hypothetical Consent and Justification. **97** (6): 313-34 '00

Realism

Wright, Crispin: Truth as Sort of Epistemic: Putnam's Peregrinations. **97** (6): 335-64 '00

Reference

García-Carpintero, Manuel: A Presuppositional Account of Reference Fixing. **97** (3): 109-47 '00

Reliabilism

Vogel, Jonathan: Reliabilism Leveled. **97** (11): 602-23 '00

Rights

Sen, Amartya: Consequential Evaluation and Practical Reasons. **97** (9): 477-502 '00

Science, Philosophy of

Eells, Ellery: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Fitelson, Branden: Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Millgram, Elijah: Coherence: The Price of the Ticket. **97** (2): 82-93 '00

Rott, Hans: Two Dogmas of Belief Revision. **97** (9): 503-22 '00

Shapiro, Lawrence A.: Multiple Realizations. **97** (12): 635-54 '00

Sober, Elliott: Evolution and the Problem of Other Minds. **97** (7): 365-86 '00

Self-Knowledge

Schick, Frederic: Surprise, Self-Knowledge, and Commonality. **97** (8): 440-53 '00

Semantics

Abbott, Barbara: Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

García-Carpintero, Manuel: A Presuppositional Account of Reference Fixing. **97** (3): 109-47 '00

Pietroski, Paul M.: On Explaining That. **97** (12): 655-62 '00

Skepticism

Vogel, Jonathan: Reliabilism Leveled. **97** (11): 602-23 '00

Testimony

Faulkner, Paul: The Social Character of Testimonial Knowledge. **97** (11): 581-601 '00

Trust

Faulkner, Paul: The Social Character of Testimonial Knowledge. **97** (11): 581-601 '00

Truth

Gaifman, Haim: What Gödel's Incompleteness Result Does and Does Not Show. **97** (8): 462-70 '00

George, Alexander: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Velleman, Daniel J.: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Wright, Crispin: Truth as Sort of Epistemic: Putnam's Peregrinations. **97** (6): 335-64 '00

Turing Machines

Copeland, B. Jack: Narrow versus Wide Mechanism: Including a Re-examination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Gaifman, Haim: What Gödel's Incompleteness Result Does and Does Not Show. **97** (8): 462-70 '00

George, Alexander: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Velleman, Daniel J.: Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Utilitarianism

Ashford, Elizabeth: Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Sen, Amartya: Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00

Williams, Bernard

Ashford, Elizabeth: Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Wittgenstein, Ludwig

Floyd, Juliet: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Putnam, Hilary: A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

INDEX OF AUTHORS

Including Contributing Authors and Reviewers

Abbott, Barbara, Fodor and Lepore on Meaning Similarity and Compositionality. **97** (8): 454-55 '00

Adler, Jonathan E., Review of Leigh S. Cauman: *First-order Logic: An Introduction.* **97** (10): 577-80 '00

Ashford, Elizabeth, Utilitarianism, Integrity, and Partiality. **97** (8): 421-39 '00

Budd, Malcolm, Review of Michael Kelly: *Encyclopedia of Aesthetics.* **97** (2): 94-99 '00

Cameron, J. R., Numbers as Types. **97** (10): 529-63 '00

Carman, Taylor, Review of William D. Blattner: *Heidegger's Temporal Idealism.* **97** (5): 308-12 '00

Cocking, Dean, Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Collins, John, Preemptive Prevention. **97** (4): 223-34 '00

Copeland, B. Jack, Narrow versus Wide Mechanism: Including a Re-examination of Turing's Views on the Mind-Machine Issue. **97** (1): 5-32 '00

Dupré, John, Review of Ian Hacking: *The Social Construction of What?* **97** (12): 673-76 '00

Eells, Ellery, Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Faulkner, Paul, The Social Character of Testimonial Knowledge. **97** (11): 581-601 '00

Fine, Arthur, Review of Nancy Cartwright: *The Dappled World: A Study of the Boundaries of Science.* By Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine. **97** (7): 403-08 '00

Fitelson, Branden, Measuring Confirmation and Evidence. By Ellery Eells and Branden Fitelson. **97** (12): 663-72 '00

Floyd, Juliet, A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Fricker, Miranda, Review of Martha Nussbaum: *Sex and Social Justice.* **97** (8): 471-75 '00

Frisch, Mathias, Review of Nancy Cartwright: *The Dappled World: A Study of the Boundaries of Science.* By Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine. **97** (7): 403-08 '00

Gaifman, Haim, What Gödel's Incompleteness Result Does and Does Not Show. **97** (8): 462-70 '00

García-Carpintero, Manuel, A Presuppositional Account of Reference Fixing. **97** (3): 109-47 '00

Gendler, Tamar Szabó, The Puzzle of Imaginative Resistance. **97** (2): 55-81 '00

George, Alexander, Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61 '00

Hall, Ned, Causation and the Price of Transitivity. **97** (4): 198-222 '00

Humphreys, Paul, Review of Wesley C. Salmon: *Causality and Explanation.* **97** (9): 523-27 '00

Kennett, Jeanette, Friendship and Moral Danger. By Dean Cocking and Jeanette Kennett. **97** (5): 278-96 '00

Kyburg, Henry E., Jr., Review of Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld: *Rethinking the Foundations of Statistics.* **97** (12): 677-80 '00

Levi, Isaac, Review of James M. Joyce: *The Foundations of Causal Decision Theory.* **97** (7): 387-402 '00

Lewis, David, Causation as Influence. **97** (4): 182-97 '00

McCauley, Robert N., Review of Paul M. Churchland and Patricia S. Church-

land: *On the Contrary: Critical Essays, 1987-1997.* **97** (5): 297-301 '00
Merikangas Darling, Karen, Review of Nancy Cartwright: *The Dappled World: A Study of the Boundaries of Science.* By Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine. **97** (7): 403-08 '00
Millgram, Elijah, Coherence: The Price of the Ticket. **97** (2): 82-93 '00

Nelkin, Dana K., Two Standpoints and the Belief in Freedom. **97** (10): 564-76 '00

Paul, L. A., Aspect Causation. **97** (4): 235-56 '00
Pettit, Philip, Two Construals of Scanlon's Contractualism. **97** (3): 148-64 '00
Pietroski, Paul M., On Explaining That. **97** (12): 655-62 '00
Pippin, Robert, Review of Michael Kelly: *Encyclopedia of Aesthetics.* **97** (2): 99-106 '00
Putnam, Hilary, A Note on Wittgenstein's "Notorious Paragraph" about the Gödel Theorem. By Juliet Floyd and Hilary Putnam. **97** (11): 624-32 '00

Rott, Hans, Two Dogmas of Belief Revision. **97** (9): 503-22 '00

Schaffer, Jonathan, Trumping Preemption. **97** (4): 165-81 '00
Schick, Frederic, Surprise, Self-Knowledge, and Commonality. **97** (8): 440-53 '00
Sen, Amartya, Consequential Evaluation and Practical Reason. **97** (9): 477-502 '00
Shapiro, Lawrence A., Multiple Realizations. **97** (12): 635-54 '00
Sober, Elliott, Evolution and the Problem of Other Minds. **97** (7): 365-86 '00
Sosa, Ernest, Review of David Lewis: *Papers in Metaphysics and Epistemology.* **97** (5): 301-07 '00
Stark, Cynthia A., Hypothetical Consent and Justification. **97** (6): 313-34 '00
Stoljar, Daniel, Physicalism and the Necessary a Posteriori. **97** (1): 33-54 '00

Tennant, Neil, Deductive versus Expressive Power: A Pre-Gödelian Predicament. **97** (5): 257-77 '00

Velleman, Daniel J., Leveling the Playing Field between Mind and Machine: A Reply to McCall. By Alexander George and Daniel J. Velleman. **97** (8): 456-61
Vogel, Jonathan, Reliabilism Leveled. **97** (11): 602-23 '00

Winsberg, Eric, Review of Nancy Cartwright: *The Dappled World: A Study of the Boundaries of Science.* By Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine. **97** (7): 403-08 '00
Wright, Crispin, Truth as Sort of Epistemic: Putnam's Peregrinations. **97** (6): 335-64 '00

INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

Blattner, William D.: *Heidegger's Temporal Idealism*—Taylor Carman. **97** (5): 308-12 '00

Cartwright, Nancy: *The Dappled World: A Study of the Boundaries of Science*—Eric Winsberg, Mathias Frisch, Karen Merikangas Darling, and Arthur Fine. **97** (7): 403-08 '00

Cauman, Leigh S.: *First-order Logic: An Introduction*—Jonathan E. Adler. **97** (10): 577-80 '00

Churchland, Patricia S., and **Paul M. Churchland:** *On the Contrary: Critical Essays, 1987-1997*—Robert N. McCauley. **97** (5): 297-301 '00

Churchland, Paul M., and **Patricia S. Churchland:** *On the Contrary: Critical Essays, 1987-1997*—Robert N. McCauley. **97** (5): 297-301 '00

Hacking, Ian: *The Social Construction of What?*—John Dupré. **97** (12): 673-76 '00

Joyce, James M.: *The Foundations of Causal Decision Theory*—Isaac Levi. **97** (7): 387-402 '00

Kadane, Joseph B., Mark J. Schervish, and **Teddy Seidenfeld:** *Rethinking the Foundations of Statistics*—Henry E. Kyburg, Jr. **97** (12): 677-80 '00

Kelly, Michael: *Encyclopedia of Aesthetics*—Malcolm Budd. **97** (2): 94-99 '00
—Robert Pippin. **97** (2): 99-106 '00

Lewis, David: *Papers in Metaphysics and Epistemology*—Ernest Sosa. **97** (5): 301-07 '00

Nussbaum, Martha: *Sex and Social Justice*—Miranda Fricker. **97** (8): 471-75 '00

Salmon, Wesley C.: *Causality and Explanation*—Paul Humphreys. **97** (9): 523-27 '00

Scanlon, T. M.: *What We Owe to Each Other*—Philip Pettit. **97** (3): 148-64 '00

Schervish, Mark J., Joseph B. Kadane, and **Teddy Seidenfeld:** *Rethinking the Foundations of Statistics*—Henry E. Kyburg, Jr. **97** (12): 677-80 '00

Seidenfeld, Teddy, Joseph B. Kadane, and **Mark Schervish:** *Rethinking the Foundations of Statistics*—Henry E. Kyburg, Jr. **97** (12): 677-80 '00

Statement of Ownership, Management, and Circulation

1 Publication Title The Journal of Philosophy	2 Publication Number 0 0 2 2 _ 3 6 2 X	3. Filing Date November 1, 2000
4 Issue Frequency Monthly	5 Number of Issues Published Annually 12	6 Annual Subscription Price $35.00 Individual $75.00 Library

7 Complete Mailing Address of Known Office of Publication *(Not printer) (Street, city, county, state, and ZIP+4)*

The Journal of Philosophy
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

Contact Person
Pam Ward

Telephone
212/666-4419

8 Complete Mailing Address of Headquarters or General Business Office of Publisher *(Not printer)*

SAME

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor *(Do not leave blank)*

Publisher *(Name and complete mailing address)*

The Journal of Philosophy
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

Editor *(Name and complete mailing address)*

Arthur Danto, President
The Journal of Philosophy
1150 Amsterdam Avenue, MC 4972 New York, New York 10027

Managing Editor *(Name and complete mailing address)*

Michael Kelly, Managing Editor
The Journal of Philosophy
1150 Amsterdam Avenue, MC 4972 New York, New York 10027

10 Owner *(Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)*

Full Name	Complete Mailing Address
THE JOURNAL OF PHILOSOPHY	1150 Amsterdam Avenue, MC 4972
	New York, New York 10027

11 Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address

12 Tax Status *(For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)*
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes.
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months *(Publisher must submit explanation of change with this statement)*

13 Publication Title	14 Issue Date for Circulation Data Below

15		Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies *(Net press run)*			4000	3925
b. Paid and/or Requested Circulation	(1)	Paid/Requested Outside-County Mail Subscriptions Stated on Form 3541 *(Include advertiser's proof and exchange copies)*	----	----
	(2)	Paid In-County Subscriptions *(Include advertiser's proof and exchange copies)*	3500	3402
	(3)	Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	----	----
	(4)	Other Classes Mailed Through the USPS	----	----
c. Total Paid and/or Requested Circulation *(Sum of 15b (1), (2),(3),and (4))* ▶			3500	3402
d. Free Distribution by Mail *(Samples, complimentary, and other free)*	(1)	Outside-County as Stated on Form 3541	150	150
	(2)	In-County as Stated on Form 3541		
	(3)	Other Classes Mailed Through the USPS		
e. Free Distribution Outside the Mail *(Carriers or other means)*				
f. Total Free Distribution *(Sum of 15d and 15e.)* ▶			150	150
g. Total Distribution *(Sum of 15c and 15f)* ▶			3650	3552
h. Copies not Distributed			350	373
i. Total *(Sum of 15g and h)* ▶			4000	3925
j. Percent Paid and/or Requested Circulation *(15c. divided by 15g. times 100)*			96	96

16 Publication of Statement of Ownership
☐ Publication required. Will be printed in the December 2000 issue of this publication. ☐ Publication not required

17. Signature and Title of Editor, Publisher, Business Manager, or Owner
Michael Kelly, Managing Editor

Date
November 1, 2000

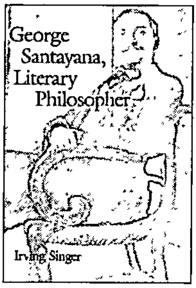
I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

Philosophy from Yale

George Santayana

Literary Philosopher

Irving Singer


George Santayana, Literary Philosopher
Irving Singer

"Firmly grounded in a comprehensively humanistic point of view, **Singer's book is rich in insight and generous in its criticism of Santayana's towering achievement** as a writer and thinker."
—Joel Porte $25.00

Humanity

A Moral History of the Twentieth Century

Jonathan Glover

"It is hard to imagine a more important book."—Peter Singer

"This is an extraordinary book: brilliant, haunting, and uniquely important."—Steven Pinker, *New York Times Book Review* $27.95

Available in paper

The Computer and the Brain

Second edition

John von Neumann

With a foreword by Paul M. Churchland and Patricia S. Churchland

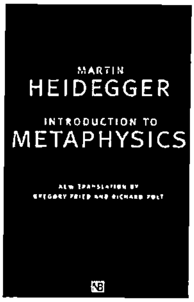
This book **represents the views of one of the greatest mathematicians of the twentieth century** on the analogies between computing machines and the living human.

$9.95 A Nota Bene Paperback

Available in paper

Introduction to Metaphysics

Martin Heidegger


MARTIN
HEIDEGGER
INTRODUCTION TO
METAPHYSICS
NEW TRANSLATION BY
GREGORY FRIED AND RICHARD POLT
NB

Translated by Gregory Fried and Richard Polt

"A superb translation combined with a nuanced understanding of the philosophical issues in the text. This is an important event within Heidegger scholarship."
—Thomas Sheehan

$14.95 A Nota Bene Paperback

Heidegger's *Polemos*

From Being to Politics

Gregory Fried

"With this book Fried clearly establishes himself as **one of the Heidegger scholars to whom one must listen."**—Robert Bernasconi

$35.00

Available in cloth & paper

A Companion to Heidegger's *Introduction to Metaphysics*

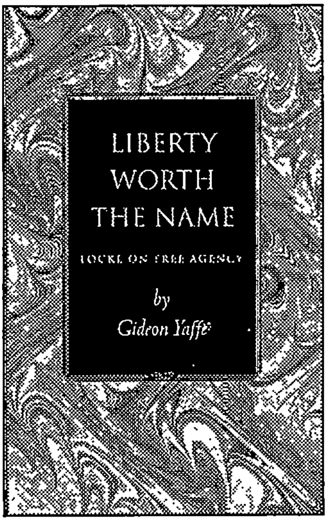
Richard Polt and Gregory Fried

"An extremely rich and insightful exploration of one of Heidegger's most important works. Anyone interested in Heidegger's thought will find rewards in this volume."
—Lawrence J. Hatab

$40.00 cloth; $18.00 paperback

Yale University Press www.yale.edu/yup/

New from Princeton



Liberty Worth the Name

Locke on Free Agency

Gideon Yaffe

This is the first comprehensive interpretation of John Locke's solution to one of philosophy's most enduring problems: free will and the nature of human agency.

"This is, quite simply, the best work on Locke's theory of agency; it is a significant contribution both to Locke studies and to the theory of action. It is clearly and elegantly written. . . ."
—Edwin McCann, University of Southern California

Princeton Monographs in Philosophy:
Harry Frankfurt, Editor
Paper $16.95 ISBN 0-691-05706-0
Cloth $49.50 ISBN 0-691-04966-1

Expanded Paperback Edition

Karl Marx's Theory of History

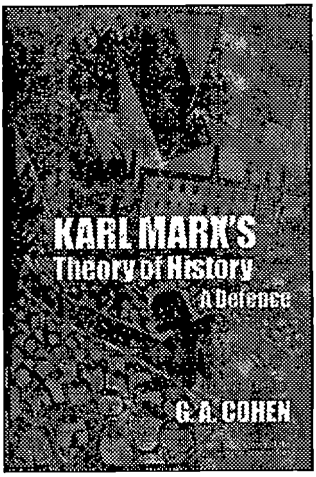
A Defence

G. A. Cohen

First published in 1978, this book rapidly established itself as a classic of modern Marxism. In the ensuing twenty years, the book has served as a flagship of a powerful intellectual movement—analytical Marxism. In this expanded edition, Cohen offers his own account of the history, and the further promise, of analytical Marxism.

"An admirable and formidable book."
—E. J. Hobsbawm

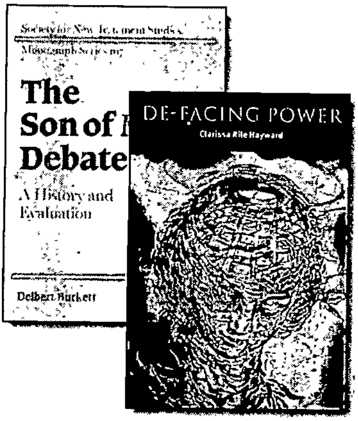
Paper $24.95 ISBN 0-691-07068-7
Due December
Not available from Princeton in the Commonwealth
(except Canada)



Princeton University Press

800-777-4726 • WWW.PUP.PRINCETON.EDU

New and Noteworthy from Cambridge



The Son of Man Debate
A History and Evaluation
Delbert Burkett
Society for New Testament Studies Monograph Series 107
0-521-66306-7 Hardback $54.95

Heidegger's Concept of Truth
Daniel O. Dahlstrom
Modern European Philosophy
0-521-64317-1 Hardback $59.95

Hegel and Aristotle
Alfredo Ferrarin
Modern European Philosophy
0-521-78314-3 Hardback $64.95

Descartes Embodied
Reading Cartesian Philosophy Through Cartesian Science
Daniel Garber
0-521-78353-4 Hardback $64.95
0-521-78973-7 Paperback $22.95

Karl Popper— The Formative Years, 1902–1945
Politics and Philosophy in Interwar Vienna
Malachi Haim Hacohen
0-521-47053-6 Hardback $54.95

De-facing Power
Clarissa Rile Hayward
Contemporary Political Theory
0-521-78079-9 Hardback $54.95
0-521-78564-2 Paperback $19.95

Plato and the Hero
Courage, Manliness and the Impersonal Good
Angela Hobbs
0-521-41733-3 Hardback $59.95

Thinking Through Philosophy
An Introduction
Chris Horner and
Emrys Westacott
0-521-62657-9 Paperback $21.95

Kant: *Critique of the Power of Judgment*
Paul Guyer and
Eric Matthews, *Editors*
Cambridge Edition of the Works of Immanuel Kant in Translation
0-521-34447-6 Hardback $64.95

Husserl and Heidegger on Human Experience
Pierre Keller
0-521-63342-7 Hardback $59.95

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

40 West 20th Street, N.Y., NY 10011-4211
Call toll-free 800-872-7423
Web site: www.cambridge.org
AmEx/MasterCard/VISA accepted. Prices subject to change.

new titles in philosophy

The
MIT
Press

Dennett's Philosophy

A Comprehensive Assessment
edited by Don Ross,
Andrew Brook,
and David Thompson

Essays that consider Dennett's work on evolution, intentionality, consciousness, ontology, ethics and free will.

A Bradford Book
448 pp. $27.95 paper

The Mind Doesn't Work That Way

The Scope and Limits of Computational Psychology
Jerry Fodor

"This book is as important as *Modularity of Mind* was almost twenty years ago."
— Jacques Mehler, Centre National de Recherche Scientifique, France, and International School for Advanced Studies, Trieste, Italy

A Bradford Book
Representation and Mind series
144 pp., 1 illus. $22.95

On the Pragmatics of Social Interaction

Preliminary Studies in the Theory of Communicative Action
Jürgen Habermas

These five lectures, from a period when Habermas was just organizing his thoughts on communication and society, offer an excellent introduction to his work, focusing on the germs of the ideas he would elaborate in *The Theory of Communicative Action.*

Studies in Contemporary German Social Thought
205 pp. $24.95

The Deleuze Connections

John Rajchman

"Anyone who thinks Deleuze is difficult should read this extraordinary book."
— D. N. Rodowick, Chair in Film Studies, King's College, London, and author of *Gilles Deleuze's Time Machine*

120 pp. $14.95 paper

Consciousness, Color, and Content

Michael Tye

"A splendid book...required reading for anyone interested in philosophical theories of consciousness, mental representation, or perception."
— Alex Byrne, Department of Linguistics and Philosophy, MIT

A Bradford Book
Representation and Mind series
184 pp., 8 illus. $26.95

Bioethics

Ancient Themes in Contemporary Issues
edited by Mark G. Kuczewski and Ronald Polansky

"A must read for those in bioethics."
— Patricia Werhane, Ruffin Professor of Ethics, University of Virginia

A Bradford Book
Basic Bioethics series
310 pp. $34.95

Metaphor in Context

Josef Stern

"A remarkable piece of work, broad in scope, artful, and a great read."
— Stephen Neale, Professor of Philosophy, Rutgers University

Studies in Contemporary German Social Thought
464 pp. $18.95 paper

Coherence in Thought and Action

Paul Thagard

"[W]ill be read by philosophers interested in almost any aspect of the subject, and more generally by anyone interested in cognitive science." — Gilbert Harman, Professor of Philosophy, Princeton University

A Bradford Book
Life and Mind: Philosophical Issues in Biology and Psychology
339 pp., 32 illus. $32.95 paper

now in paperback

Oratio Obliqua, Oratio Recta

An Essay on Metarepresentation
François Recanati

"[O]ffers a distinctive and original perspective on central problems in the philosophy of language."
— Robert Stalnaker, Department of Linguistics and Philosophy, MIT

A Bradford Book
Representation and Mind series
450 pp., 2 illus. $24.95 paper

A Zone Book

Detour and Access

Strategies of Meaning in China and Greece
François Jullien
translated by Sophie Hawkes

An exploration of indirect and allusive modes of expression in ancient Chinese culture.

Distributed for Zone Books
435 pp. $32

Sublime Understanding

Aesthetic Reflection in Kant and Hegel
Kirk Pillow

"[O]ne of the most useful books on Kant's aesthetics to have appeared in a decade, and the most fruitful juxtaposition of Kantian and Hegelian aesthetics to have appeared in a generation."
— Paul Guyer, Murray Professor in the Humanities, University of Pennsylvania

Studies in Contemporary German Social Thought
386 pp. $37.95

Beyond Hegel and Nietzsche

Philosophy, Culture, and Agency
Elliot L. Jurist

This book places Hegel and Nietzsche in conversation with each other, reassessing their relationship in a way that affirms its complexity.

Studies in Contemporary German Social Thought
394pp. $37.95

German Realists in the Nineteenth Century

Georg Lukåks
translated by Jeremy Gaines and Paul Keast
edited and with an introduction and notes by Rodney Livingstone

Literary studies on major German writers of the 19th century by one of the most controversial Marxist philosophers of the 20th century.

396 pp. $24.95 paper

Endgames

The Irreconcilable Nature of Modernity
Essays and Lectures
Albrecht Wellmer
translated by David Midgley

Includes essays on political philosophy, the continuing legacy of the Frankfurt school, architecture, and a number of Wellmer's dialogue partners, from Adorno to Arendt.

Studies in Contemporary German Social Thought
368 pp. $24.95 paper

Reality Transformed

Film as Meaning and Technique
Irving Singer

"Taking as his premise the currently unfashionable idea that narrative cinema has something to do with life, Singer demonstrates how it transforms our common experience into art."
— *Sight & Sound*

232 pp., 3 illus. $16.95 paper

A Zone Book

A Thousand Years of Nonlinear History

Manuel De Landa

"Forcefully challenges habituated understandings of 'history', 'urban' and 'economics'." — Christopher Hight, *AA Files*

Distributed for Zone Books
333 pp., 10 illus. $16 paper

To order call **800-356-0343** (US & Canada)
or 617-625-8569. Prices subject to change without notice.

http://mitpress.mit.edu

Social Thoughts from CALIFORNIA

Weimar

A Jurisprudence of Crisis

Arthur Jacobson and Bernhard Schlink, Editors

"Includes the most significant and influential texts representative of the political and conceptual diversity of the intellectual approaches of that time.... Very significant for contemporary debates about the relationship between state, law, and constitution." —Ulrich Karl Preuss, Freie Universität Berlin

Philosophy, Social Theory, and the Rule of Law, $55.00 cloth

Walter Benjamin's Other History

Of Stones, Animals, Human Beings, and Angels

Beatrice Hanssen

New in Paperback—"Hanssen's exacting, expansive study of the ways Benjamin reconceives history and nature in one another's presence, or distance, is part of the increasing recognition of what it must take intellectually and imaginatively to come to terms with this thinker's soaring innovations."

—Stanley Cavell, Harvard University

Weimar and Now: German Cultural Criticism, $16.95 paper

Songs without Music

Aesthetic Dimensions of Law and Justice

Desmond Manderson

Creatively using music as a model, Manderson demonstrates that law is not a sterile, rational structure, but a cultural form to be valued and enhanced through rhetoric and metaphors, form, images, and symbols. It successfully traverses several disciplines and builds an original and persuasive argument for a legal aesthetic.

Philosophy, Social Theory, and the Rule of Law, $55.00 cloth

In the Shadow of Catastrophe

German Intellectuals Between Apocalypse and Enlightenment

Anson Rabinbach

New in Paperback—"We are largely in the dark about the connections between revolutionary visions and reactionary brutality, but Rabinbach invites us to confront these matters more candidly in relation to texts from the fascist era that are rarely considered together."—David S. Luft, *Central European History*

Weimar and Now: German Cultural Criticism, $18.95 paper

At bookstores or order (800) 822-6657 • www.ucpress.edu

UNIVERSITY OF CALIFORNIA PRESS